UTEC-CSc-79-039

Semi-Annual Technical Report

NOISE SUPPRESSION METHODS FOR ROBUST SPEECH PROCESSING

Contractor: University of Utah
Effective Date: 2 January 1979
Expiration Date: 30 September 1980
Reporting Period: 1 October 1978 - 31 March 1979

Principal Investigator: Dr. Steven F. Boll
Telephone: (801) 581-8224

Sponsored by
Defense Advanced Research Projects Agency (DoD)
ARPA Order No. 3301
Monitored by Naval Research Laboratory
Under Contract No. N00173-79-C-0045

April 1979

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

14 ~~1. REPORT NUMBER~~
UTEC-CSc-79-039

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

6 ~~4. TITLE (and Subtitle)~~
Noise Suppression Methods for Robust Speech Processing.

9 ~~5. TYPE OF REPORT & PERIOD COVERED~~
Semi-Annual Technical rept.
1 Oct [redacted]78-31 Mar [redacted]79,

~~6. PERFORMING ORG. REPORT NUMBER~~

10 ~~7. AUTHOR(s)~~
Dr. Steven F. Boll, Dennis Pulsipher, William Done, Ben Cox [redacted] K. Rushforth ~~[illegible]~~ ~~[illegible]~~
Craig

8. CONTRACT OR GRANT NUMBER(s)
15 N00173-79-C-0045, Jew
V/ARPA Order-3301

9. PERFORMING ORGANIZATION NAME AND ADDRESS
University of Utah
Computer Science Department
Salt Lake City, Utah 84112

~~10. PROGRAM ELEMENT, PROJECT, TASK~~ AREA & WORK UNIT NUMBERS
Project: 76-RPA-3301

11. CONTROLLING OFFICE NAME AND ADDRESS
Defense Advanced Research Project Agency (DoD)
1400 Wilson Boulevard
Washington, D.C. 22209

~~12. REPORT DATE~~
11 Apr[redacted]79

13. NUMBER OF PAGES
58

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)
Naval Research Laboratory
4555 Overlook Avenue, S.W.
Mail Code 2415-A.M.

12 59p.

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Digital noise suppression; Linear Predictive Coding; Narrow band coded speech; Adaptive noise cancellation; Weiner filtering; Power spectrum; Autocorrelation, Spectral Averaging for Bias Estimation and Removal (SABER); Widrow-Hoff LMS Algorithm; Pole-zero modeling estimation; Constant-Q Transform; Non-Parametric speech activity detector, Pitch/rate speech modifications.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Robust speech processing in practical operating environments requires effective environmental and processor noise suppression. This report describes the technical findings and accomplishments ~~during this reporting period for the research program funded~~ to develop real time, compressed speech analysis-synthesis algorithms whose performance is invariant under signal contamination. Fulfillment of this requirement is necessary to insure reliable secure compressed speech transmission within realistic

over

404 949

20. ABSTRACT CON'T

military command and control environments. Overall contributions resulting from this research program include the understanding of how environmental noise degrades narrow band, coded speech, development of appropriate real time noise suppression algorithms, and development of speech parameter identification methods that consider signal contamination as a fundamental element in the estimation process. This report describes ~~the current research and~~ results in the areas of noise suppression using the spectral subtraction algorithm, dual input adaptive noise cancellation using the LMS algorithm, pole-zero parameter estimation, nonparametric-rank order statistics with applications to Robust Speech activity detection, spectral analysis and synthesis using the constant-Q transform, and pitch and rate changes to speech using the constant-Q transform.

TABLE OF CONTENTS

Page

I. D D Form 1473 i

II. List of Figures iii

III. REPORT SUMMARY

Section I Summary Program for Reporting Period 1

IV. ABSTRACTS

A Spectral Subtraction Algorithm For Suppression of Acoustic Noise in Speech 2
Steven F. Boll

Reduction of Nonstationary Noise In Speech Using LMS Adaptive Noise Cancelling 3
D. Pulsipher, S. Boll, C. Rushforth L. Timothy

Rank-Order Speech Classification Algorithm (RASCAL) 4
Ben Cox, L. Timothy

Estimating the Parameters of a Noisy All-Pole Process Using Pole-Zero Modeling 5
W.J. Done, C.K. Rushforth

Evaluation of the Steiglitz Algorithm for Estimating the Parameters of an ARMA Process 7
W.J. Done, C.K. Rushforth

Rate/Pitch Modification Using the Constant-Q Transform 9
James E. Youngberg

TABLE OF CONTENTS

Page

V. RESEARCH ACTIVITIES

A Spectral Subtraction Algorithm For Suppression of Acoustic Noise in Speech 10

Steven F. Boll

Reduction of Nonstationary Noise In Speech Using LMS Adaptive Noise Cancelling 30

D. Pulsipher, S. Boll, C. Rushforth . . . .
L. Timothy

Rank-Order Speech Classification Algorithm (RASCAL) 35

Ben Cox, L.K. Timothy

Estimating the Parameters of a Noisy All-Pole Process Using Pole-Zero Modeling 41

W.J. Done, C.K. Rushforth

Evaluation of the Steiglitz Algorithm for Estimating the Parameters of an ARMA Process 46

W.J. Done, C.K. Rushforth

Rate/Pitch Modification Using the Constant-Q Transform 51

James E. Youngberg

LIST OF FIGURES

Figure	Page
Noise Only Signal, Subtractive Filters	25
Noise Only Signal, Magnitude Spectra	25
Voiced Speech, Subtractive Filters	26
Voiced Speech, Magnitude Spectra	26
System Block Diagram	27
Filter Length vs. Noise Reduction in dB	31
Noise Cancelling Model	32
Basic Data Generation Model	32
Realistic Data Generation Model	32
Equivalent Data Generation Model	32
Block Diagram of Signal Classification Method	37
Partitioning of Speech Spectrum unto Four Contiguous Bands that Contribute Equally to Articulation Index with Frequency Ranges of 200 to 3200 Hz.	37
20-Sample Recognition Rate for the Simultaneous Decision Procedure	38
Comparison of Four Estimators of a(1).	43
NR and LPC Estimates of a(1) with $\pm\sigma$ limits	43
10-Pole, 2 Zero Spectrum to be Indentified	47
System Output When excited by Impulse	47
Estimate of Model Spectrum from Impulse-Excited Output	47
System Output when Excited by Impulse Train	47

LIST OF FIGURES Continued

Figure	Page
Modified System Output After Cepstral Processing	47
Estimate of Model Spectrum from Modified Impulse-train-excited Output	47
The Relationship of Axis Scaling in the Time and Spectral Domains	52
A Single Channel of Harmonic Scaler	52

Section I

Summary of Program for

Reporting Period

Program Objectives

To develop practical, low cost, real time methods for suppressing noise which has been acoustically added to speech.

To demonstrate that through the incorporation of the noise suppression methods, speech can be effectively analysed for narrow band digital transmission in practical operating environments.

Summary of Tasks and Results

Introduction

This Semi-Annual technical report describes the status of work performed during the period 1 October 1978 through 31 March 1979 under ARPA order 3301, contract N00173-79-C-0045 with Naval Research Laboratories.

A SPECTRAL SUBTRACTION ALGORITHM FOR SUPPRESSION OF ACOUSTIC NOISE IN SPEECH

Steven F. Boll

Abstract

Spectral subtraction has been shown to be an effective approach for reducing ambient acoustic noise in order to improve the intelligibility and quality of digitally compressed speech. This paper presents a set of implementation specifications to improve algorithm performance and minimize algorithm computation and memory requirements. It is shown spectral subtraction can be implemented in terms of a nonstationary, multiplicative, frequency domain filter which changes with the time varying spectral characteristics of the speech. Using this filter a speech activity detector is defined and used to allow the algorithm to adapt automatically to changing ambient noise environments. Also the bandwidth information of this filter is used to further reduce the residual narrowband noise components which remain after spectral subtraction.

REDUCTION OF NONSTATIONARY NOISE IN SPEECH USING LMS ADAPTIVE NOISE CANCELLING

Dennis Pulsipher, Steven F. Boll, Craig Rushforth, LaMar Timothy

Abstract

Nonstationary acoustic noise with energy possibly equal to or greater than the speech is suppressed using a two microphone implementation of adaptive noise cancellation. The primary noise added to the speech is reduced by subtracting a filtered version of the second microphone reference noise. The reference noise filter is adaptively up dated using the Widrow-Hoff LMS algorithm. The effectiveness of noise suppression depends directly on the ability of the filter to estimate the transfer function relating the primary and reference noise channels. A study of the filter length required to achieve a desired noise reduction level in a hard-walled room is presented. Results demonstrating noise reduction in excess 10dB in an environment with 0dB signal to noise ratio are presented.

This abstract is taken from the Ph.D dissertation of Dennis Pulsipher. This dissertation entitled "Application of Adaptive Noise Cancellationto Noise Reduction in Audio Signals" has been published as a technical report No. UTEC-CSc-79-022.

RANK-ORDER SPEECH CLASSIFICATION ALGORITHM
(RASCAL)

Ben Cox

L. K. Timothy

Abstract

This paper describes a theoretical and experimental investigation for detecting the presence of speech in wide band noise. A robust algorithm for making the silence-voiced-unvoiced decision is described. This algorithm is based on a nonparametric rank-order statistical signal-detection scheme that does not require a training set of data and maintains a constant false-alarm rate for a broad class of noise inputs corresponding to a single decision threshold. The nonparametric rank-order decision procedure is the multiple use of the two-sample Savage T statistic. The performance of this detector is evaluated and compared to that obtained by manually classifying twenty recorded utterances with 39, 30, 20, 10, and 0 decibel signal-to-noise ratios. In limited testing, the average probability of misclassification is less that 5 percent, 12 percent, and 55 percent for signal-to-noise ratios of 39, 20, and 0 decibels respectively.

ESTIMATING THE PARAMETERS OF A NOISY ALL-POLE PROCESS USING POLE-ZERO MODELING

W. J. DONE

C. K. RUSHFORTH

Abstract

Linear predictive coding (LPC) has been successfully applied to the encoding of speech and other time series. It has been widely observed, however, that the performance of an LPC algorithm deteriorates rapidly in the presence of background noise. In this paper, we describe and discuss one approach to the identification of a time series corrupted by additive white noise.

A common approach to this problem is to prefilter the noisy time series, and then to apply an estimation algorithm which treats the time series as if it were noise-free. We describe an alternative approach which involves modifying the time-series model at the outset to the account for the presence of noise. An estimation algorithm is then developed for this modified model. We discuss the development of the model, the estimation algorithm, and some representative experimental results.

This abstract is taken from the Ph.D. dissertation of W.J. Done entilted, "Estimation of the Parameters of an Autoregressive Process in the Presence of Additive White Noise." This dissertation has been published as technical report No. UTEC-CSc-79-021.

EVALUATION OF THE STEIGLITZ ALGORITHM FOR ESTIMATING THE PARAMETERS OF AN ARMA PROCESS

W. J. Done
C. K. Rushforth

Abstract

Steiglitz has recently described an algorithm for estimating the parameters of an autoregressive-moving-average (ARMA) process. This algorithm has application, for example, to the problem of determining the poles and zeros of the vocal-tract transfer function.

In this paper, we report and discuss the results of a number of simulations conducted using the Steiglitz algorithm. The bulk of the experiments involved driving the ten-pole, two-zero filter described in (2) with a single pulse, with a short pulse train, and with samples of white Gaussian noise. In each of these cases, we evaluated the effects of such processing options as windowing, preemphasis, and cepstral-domain filtering. We also discuss and compare results obtained by applying the Steiglitz algorithm and a Newton-Raphson conditional maximum-likelihood algorithm to a first-order process.

This abstract is taken from the Ph.D. dissertation of W.J. Done entitled, "Estimation of the Parameters of an Autoregressive Process in the Presence of Additive White Noise." This dissertation has been published as technical report, No. UTEC-CSc-79-021.

RATE/PITCH MODIFICATION
USING THE CONSTANT-Q TRANSFORM

James E. Youngberg

Abstract

Modification of the rate of occurrence of acoustic events without altering frequency content, and modification of pitch without changing time scale are presented as equivalent problems. While the short-time Fourier transform has been used to solve the rate modification problem, it is not a natural tool. It lacks the scaling property of the Fourier transform. The Constant-Q transform, on the other hand, exhibits this properly. A more natural rate/pitch modification system using the Constant-Q transform is presented which performs well with rate/pitch changes by factors of between one-third and three.

A SPECTRAL SUBTRACTION ALGORITHM FOR SUPPRESSION OF ACOUSTIC NOISE IN SPEECH

Steven F. Boll

To be presented at ICASSP-79
April 2-4, 1979
Washington, D.C.

A Spectral Subtraction Algorithm for Suppression of Acoustic Noise in Speech

Steven F. Boll
Computer Science Department
University of Utah
Salt Lake City, Utah 84112

Spectral subtraction has been shown to be an effective approach for reducing ambient acoustic noise in order to improve the intelligibility and quality of digitally compressed speech. This paper presents a set of implementation specifications to improve algorithm performance and minimize algorithm computation and memory requirements. It is shown spectral subtraction can be implemented in terms of a nonstationary, multiplicative, frequency domain filter which changes with the time varying spectral characteristics of the speech. Using this filter a speech activity detector is defined and used to allow the algorithm to adapt automatically to changing ambient noise environments. Also the bandwidth information of this filter is used to further reduce the residual narrowband noise components which remain after spectral subtraction.

Introduction

Digital speech compression systems operating in environments with high ambient acoustic noise may require additional noise suppression to process speech having acceptable intelligibility and quality [1]. Previous results to suppress noise using the spectral subtraction approach have demonstrated quantitative improvements in quality and intelligibility [2], [3]. This paper describes a number of techniques for improving the efficiency and effectiveness of this approach. It is shown that the algorithm can be implemented in terms of a nonstationary, multiplicative, frequency domain filter. Characteristics of this filter provide information for further reduction of spectral error and detection of speech activity. In addition techniques are presented for increasing algorithm efficiency, decreasing memory requirements, decreasing processing delay, and simplifying requirements for interfacing the noise suppressor with the subsequent speech compression analyzer.

Signal Estimation Using Spectral Subtraction

Signal x(i) digitized from a single microphone consists of the sum of speech Sp(i) and ambient acoustic noise n(i). It is assumed that the noise is locally stationary to the extent that average value of its spectral magnitude during speech activity is equal to that measured just prior to speech activity. Using these assumptions the spectral subtraction algorithm attempts to suppress the additive acoustic noise component n(i) from x(i) by the following steps:

1. Segment the noisy data into windowed analysis blocks of length M samples, $x(i), i=0,1...,M-1$.

2. Compute the N point DFT X(k) of data x(i).

3. Estimate the speech spectrum S(k) by subtracting the average noise spectral magnitude, $B(k) = \text{ave}|N(k)|$, calculated during non-speech activity, from $|X(k)|$:

$$S(k) = [|X(k)|-B(k)] \exp(j\ \text{ARG}[X(k)]) \quad k=0,1,...,N-1$$

The motivation behind this approach is to subtract from the noisy speech spectrum, an estimate of the noise spectrum which is readily available. The magnitude of N(k) is replaced by its average value, B(k), and the phase of N(k) is replaced by the phase of X(k).

The spectral error using this approach is given by

$$S(k)-Sp(k) = N(k)-B(k) \exp(j\ \text{ARG}[X(k)])$$

A simple method for reducing this error is half-wave rectification. With it the estimator becomes

$$S(k) = \begin{cases} \{X(k)-B(k)\}\exp(j\ \mathrm{ARG}[X(k)]) & |X(k)|>B(k) \\ 0 & |X(k)|<B(k) \end{cases}$$

Multiplicative Filter

The spectral subtraction estimator can be compactly defined using a multiplicative frequency filter, $H(k)$:

$$H(k) = (1-B(k)/|X(k)|)(1/2 + 1/2\ \mathrm{SGN}(|X(k)|-B(k)))$$

The speech estimate $S(k)$ is then given by $S(k)=H(k)X(k)$. Examination of the expression for $H(k)$ shows that $H(k) = 0$ when $|X(k)|<B(k)$, (band stop) and $H(k)\sim 1$ for $|X(k)|>>B(k)$, (band pass). In addition an estimate of the signal to noise ratio SNR is directly available from $H(k)$ at each frequency bin k:

$$\mathrm{SNR}(k) = S(k)/B(k) = H(k)/(1-H(k))$$

Residual Noise Suppression

After half-wave rectification speech plus noise above $B(k)$ remains. In the absence of speech activity, the noise residual $N(k)-B(k)\ \exp\ (j\ \mathrm{ARG}[n(k)])$ will exhibit itself as randomly spaced narrowband spikes separated by intervals, having zero magnitude. The corresponding frequency filters $H(k)$ will have the same zero magnitude intervals. Non-zero amplitudes will have values given by

$$H(k) = 1-B(k)/|N(k)|$$

These values, being deviations of the noise magnitude spectrum above its mean correspond to the noise residual. Assuming the noise to be a zero mean, Gaussian process, the magnitude spectrum of |N| will have a Rayleigh distribution. Using this information it can be shown that less than 1% of the time will H(k) exceed a value of 0.6 (2.5 times its mean, B(k)) when speech is absent. This suggests that the noise residual could be eliminated 99% of the time by simply zeroing all spectral components which corresponds to values of H(k) less than 0.6. However, during speech activity, assuming Gaussian speech and a signal to noise ratio of 10dB, H(k) will take on values below 0.6, about 36% of the time. Thus simply rejecting all spectra X(k) corresponding to H(k) below 0.6 could in some instances incorrectly remove low energy speech spectra.

In order to reduce the noise residual but retain low energy speech in X(k), a magnitude plus bandwidth measurement test is used. Sections of H(k) having bandwidths less than 300Hz and amplitudes less than 0.6 are classified as being due only to noise. Here bandwidth is defined as the distance between successive frequency bins having zero amplitude. The 300Hz figure was empirically determined after examining an ensemble of subtractive filter frequency responses taken during non-speech activity using helicopter noise. These noise only sections are attenuated by an additional 20dB.

This secondary noise suppression procedure was applied to all values of H(k) above 800Hz. Below 800Hz narrowband harmonics essential to accurate pitch detection can be present. This procedure could incorrectly attenuate them causing pitch tracking errors. Therefore in this frequency region only bias removal and half-wave rectification is employed. The 800Hz value was picked to equal the cutoff frequency of the low-pass filter applied to the signal prior to down sampling for SIFT [4] pitch detection. Figure 1 shows examples of subtractive filters and corresponding magnitude spectra before and after residual noise reduction computed for a frame of noise only signal. Figure 2 shows examples of subtractive filters and corresponding magnitude spectra before and after residual noise reduction during voiced speech.

Algorithm Implementation

The task of spectral subtraction is to provide the vocoder analyzer with a buffer of noise suppressed speech in a time interval which is not only less than the buffer length time but which is also short enough to allow the analyzer to compute and transmit the vocoder channel parameters. This interfacing constraint imposes certain conditions on the implementation. The algorithm should use the same buffer size as the analyzer. Assuming a single processor it must compute the noise suppressed speech in the time left over after the analyzer calculations. It must supply the processed speech with minimum delay. In addition to the basic noise suppression procedures, it must monitor the signal to noise environment and update the average noise bias spectrum B(k) if necessary.

Data Segmentation

Buffer lengths of speech compression analyzers come in all sizes. Matching the noise suppression analysis buffer to that used by the vocoder results in the simplest implementation. This approach, however, leads to two operational compromises. First, if the buffer is not a power of two then zeros must be appended before transforming. Second, if buffer lengths are to be matched, with minimum delay, then no overlapping (and thus no windowing) is allowed. The effect of padding with zeros simply means lower efficiency (fewer points processed per FFT). It has a positive effect of reducing the amount of temporal aliasing due to spectral modification [5]. No overlap of time windows doubles the processing speed. The possible detrimental effect of having no time window consists of inducing discontinuities at the buffer boundaries. Reconstituted waveforms from successive analysis buffers will not necessarily agree at the boundary. In fact, in listening to the processed speech entering the vocoder, a low-level but distinct clicking sound can sometimes be heard having a frequency equal to the analysis frame rate. The clicking is due to waveform discontinuities at the boundaries. If the data had been weighted by half-overlapped hanning windows, the discontinuity effect could be minimized. However, since the speech is to be further processed by a compression analyzer using the same buffer size, the discontinuities do not cause noticeable problems.

Bidirectional Biplexed DFT

Spectral subtraction requires two DFT's to be performed: a forward transform of the noisy signal x(i) and an inverse transform of the noise suppressed spectrum, S(k)=X(k)H(k). Armantrout [6] developed a biplexed DFT which simultaneously computes the forward transform of x(i) and the inverse transform of S(k) from the previous frame. The loading procedure is given as

$$RE(i) = xo(i) + SR(i)/N$$

$$Im(i) = xe(i) - SI(i)/N$$

where $xe(i) = (x(i) + x(N-i))/2$, even part of $x(i)$

$xo(i) = (x(i) - x(N-i))/2$, odd part of $x(i)$

$SR(i)$ = Real part of $S(i)$

$SI(i)$ = Imaginary part of $S(i)$

N = DFT size

Let $C(k) + jD(k) = \text{DFT}\{RE(i)+jIM(i)\}$

Then

$$s(k) = C(k)$$

$Re\{X(k)\} = (D(k) + D(N-k))/2$, even part of $D(k)$

$Im\{X(k)\} = (D(k) + D(N-k))/2$, odd part of $D(k)$

where

$s(k)$ equals the inverse DFT of $S(k)$

$Re\{X(k)\}$ = Real Part of $X(k)$

$Im\{X(k)\}$ = Imaginary part of $X(k)$

In addition, the even-odd symmetries of the signals can be used to reduce the storage requirement in half. That is, the even part of the signal can be stored in the first N/2+1 locations and the odd part of the signal in the last N/2-1 locations.

Speech Activity Detection

Effective noise suppression requires an accurate estimate of the average noise bias, B(k). If the ambient noise becomes either louder or softer, the bias should be updated during the next interval of non-speech activity.

For detecting the absence of speech activity during a stationary noise interval and/or detecting a decrease in the noise bias, the estimated signal to noise ratio:

$$SNR(k)=H(k)/(1-H(k))=S(k)/B(k)$$

can be used. Computing the average SNR(k) over all frequency bins provides a measure the relative energy of S to B. During the absence of speech activity, the SNR was found to be less than -12dB over a wide range of noise environments. This measure also can detect when the ambient noise becomes less. In this instance more values of X(k) will lie below B(k)and thus more values of H(k) will be zero driving the average value down. Thus the measure H/(1-H) averaged over all frequency bins compared with the threshold -12dB was used to signal speech absence and/or noise bias reduction.

Noise Bias Increase Detection

Detecting when the average noise bias has become louder presents a more difficult problem since spectra above the noise mean is assumed to be speech. As the noise increases a larger percentage of $X(k)$ lies above $B(k)$. Thus if $N(k) \gg B(k)$ then $H(k) \sim 1$. This unfortunately is the identical situation found during a high signal to noise ratio environment. The measure that is needed is $N(k)/B(k)$ or equivalently $X(k)/B(k)$ for $Sp(k)=0$. A procedure used to obtain $N(k)/B(k)$ was to average $X(k)/B(k) = 1/1-H(k))$over the top 300Hz of the base band. If this average was greater than 10dB for ten consecutive analysis frames then the noise bias is updated.

Automatic Operation

Using the speech activity and bias increase detectors the spectral subtraction algorithm will run without operator intervention. The detectors provide one of many possible schemes for adaptive operation in a changing noise environment. Others are possible and proper procedures for correct adaption still remain a research issue probably best resolved using a real-time system employed in actual operating environments.

A block diagram showing the various algorithm procedures is given in Figure 3.

Discussion

Omitting the windowing and half overlapping simplifies the interface requirements with the follow on vocoder and doubles the throughput per transform of the algorithm. This approach induces discontinuities at the boundaries which are essentially ignored by the speech analyzer. Using the bidirectional, biplexed DFT produces only one frame of delay and takes advantage of the symmetries of the real data to reduce FFT computation by about one-half. Reduction of the residual noise left after subtraction using the amplitude-bandwidth test removes the majority of the noise residual while retaining wide bandwidth, low energy speech. However, noise spectral components which exceed 2.5 times its mean or with bandwidths greater than 300Hz will remain. These components, due both to statistical randomness and nonstationary, remain due to their resemblance to speech spectra. Thus the algorithm is biased towards keeping low energy speech and high energy noise.

A final modification to the multiplicative filter to suppress the acoustic effect of the remaining noise is to replace the zero amplitude frequency bins in H with a small constant. Using 0.1 instead of 0.0 brings the noise floor up, insures that the magnitude spectrum is now everywhere positive, and reinstates the natural ambient noise environment only now attenuated by approximately 20dB.

It should be apparent that as the signal and noise energies become equal or the noise becomes highly nonstationary this algorithm will break down. Speech intelligibility in these situations can be improved using noise suppression microphones [1] and/or two microphone adaptive noise cancellation procedures [7].

References

1. C. Teacher and H. Watkins, ANDVT Microphone and Audio System Study, Final Report, Ketran Inc., No. 1159, Aug. 1978.

2. S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Transactions on Acoustics Speech and Signal Processing, to Appear.

3. S. F. Boll, "Suppression of Noise in Speech Using the SABER Method", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Tulsa, OK, pp 606-609, April, 1978.

4. J.D. Markel and A.H. Gray, Linear Prediction of Speech, Springer-Verla, New York, New York, 1976.

5. J. Allen, "Short Term Spectral Analysis, Synthesis, and Modification by the Discrete Fourier Transform," IEEE Trans on Acoustics, Speech and Signal Processing, Vol ASSP-25, p. 235-239, June, 1977.

6. R. Armantrout, Private communication.

7. D. Pulsipher, S.F. Boll, C. Rushforth, L.K. Timothy, "Reduction of Nonstationary Acoustic Noise in Speech Using LMS Adaptive Noise Cancelling", Proceedings of the 1979 International Conference on Acoustics Speech and Signal Processing, Wash. D.C., April 1979

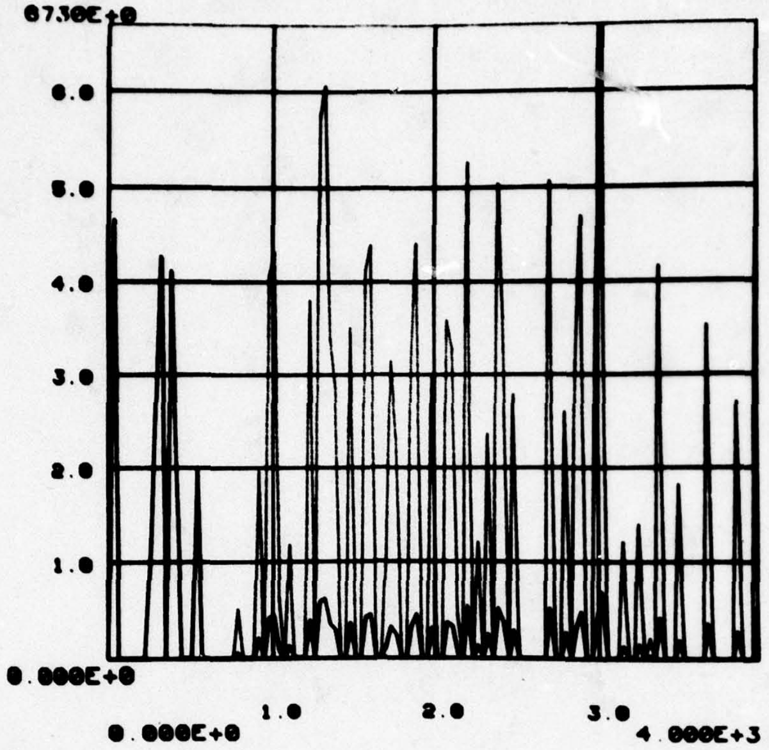


Figure 1a: Noise only signal, subtractive filters

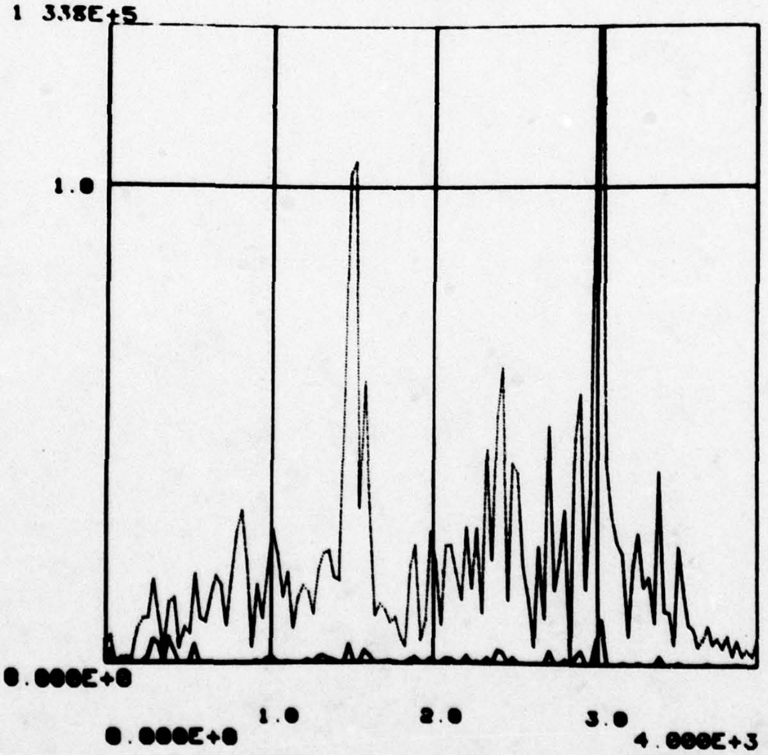


Figure 1b: Noise only signal, magnitude spectra.

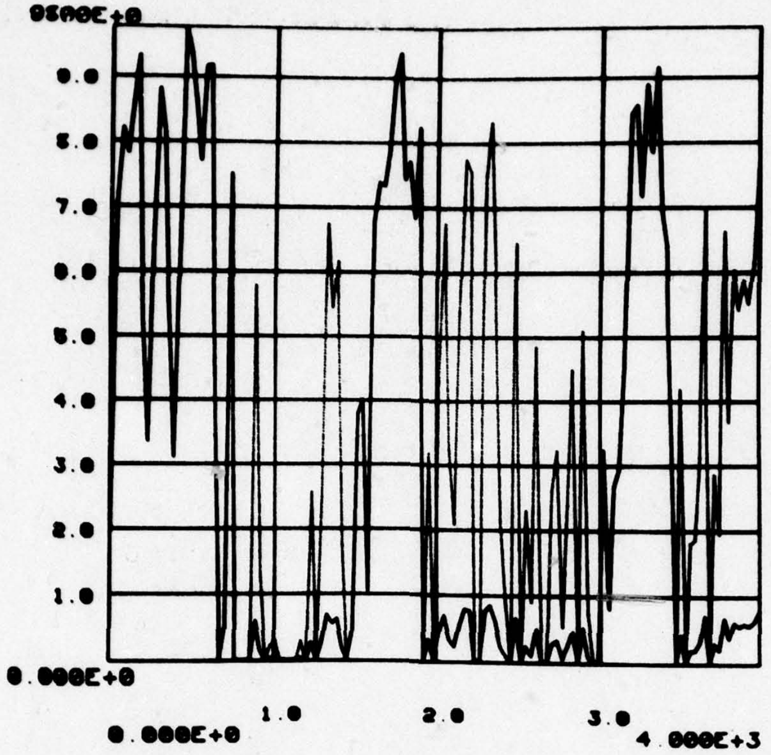


Figure 2a: Voiced Speech, Subtractive Filters.

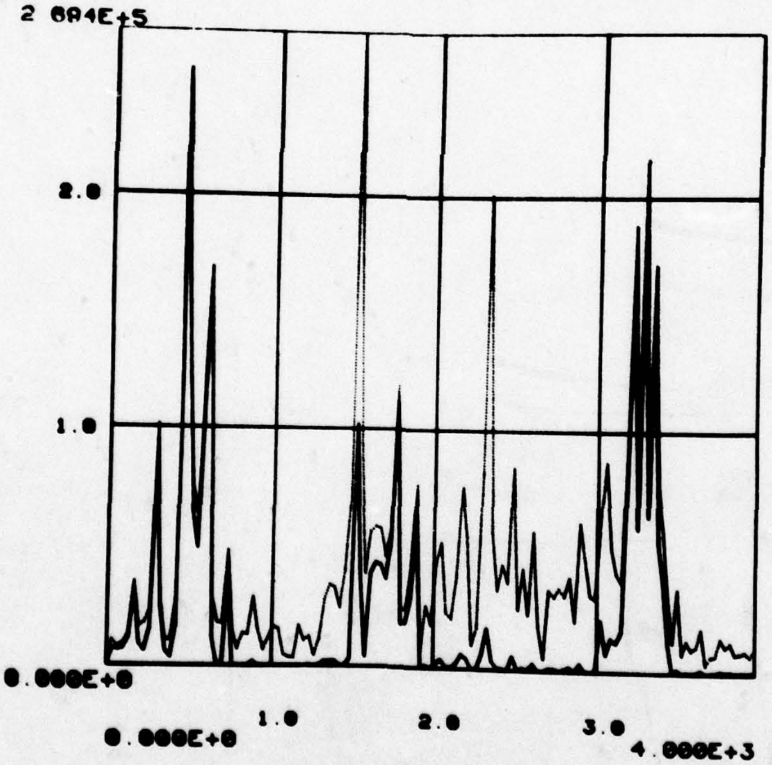


Figure 2b: Voiced Speech, Magnitude Spectra.

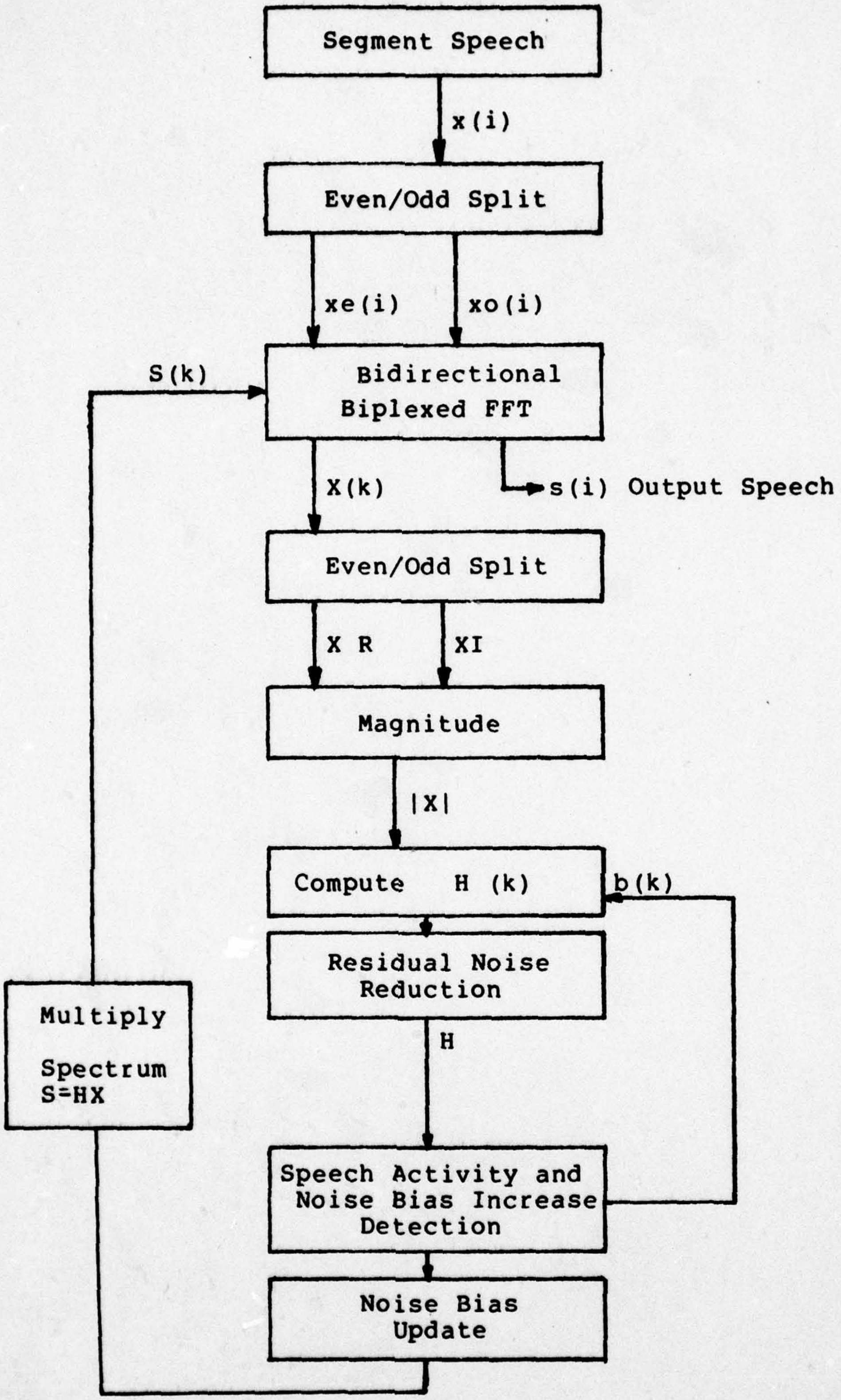


Figure 3
System Block Diagram

REDUCTION OF NONSTATIONARY ACOUSTIC NOISE IN SPEECH USING LMS ADAPTIVE NOISE CANCELLING

Steven F. Boll
Computer Science Dpt.
University of Utah
Salt Lake City, UT

Craig Rushforth
E.E. Dept.
University of Utah
Salt Lake City, UT

Dennis Pulsipher
Sandia Laboratory
Livermore, CA

LaMar Timothy
E.E. Dept.
Unversity of Utah
Salt Lake City, UT

December 1978

To be presented at ICASSP-79

April 2-4, 1979

Washington, D.C.

REDUCTION OF NONSTATIONARY ACOUSTIC NOISE IN SPEECH USING LMS ADAPTIVE NOISE CANCELLING

Dennis Pulsipher Steven F. Boll Craig Rushforth LaMar Timothy

Sandia Laboratory University of Utah University of Utah University of Utah

Nonstationary acoustic noise with energy possibly equal to or greater than the speech is suppressed using a two microphone implementation of adaptive noise cancellation. The primary noise added to the speech is reduced by subtracting a filtered version of the second microphone reference noise. The reference noise filter is adaptively updated using the Widrow-Hoff LMS algorithm [1]. The effectiveness of noise suppression depends directly on the ability of the filter to estimate the transfer function relating the primary and reference noise channels. A study of the filter length required to achieve a desired noise reduction level in a hard-walled room is presented. Results demonstrating noise reduction in excess 10dB in an environment with 0dB signal noise ratio are presented.

Introduction

Let us assume that we are given x(t), the sum of two mutually uncorrelated signals, s(t) and n(t), and a third signal v(t), which is mutually uncorrelated with s(t). We can then form a signal estimate

(1) $\bar{s}(t) = x(t)-u(t)=s(t)+[n(t)-u(t)]$

where u(t) is a noise estimate which we will constrain to be a linearly filtered version of v(t), (see Figure 1). Minimizing the mean output power causes the signal estimate $\bar{s}(t)$ to be a least mean squares fit to the signal s(t). The minimization, of course, must be carried out by choosing an h(t) (the impulse response of the filter through which v(t) is passed to generate u(t)) which minimizes the power in $\bar{s}(t)$. We, then, are looking for h(t) which satisfies:

$$\underset{h(t)}{\mathrm{Min}}[E\{\bar{s}(t)^2\}]$$

Block Solution

Let v_n, x_n, s_n, etc. be the value of the corresponding signal at time nT, where T is the sampling interval.
Define the vectors

(2) $V_n=[v(n)\dots v(n-L+1)]$ $H_n=[h(1,n)\dots h(L,n)]$

where L is the length of the filter to be estimated and H is the filter.
Defining

(3) $P = E\{x_n V_n\}$ $R = E\{V_n V_n^T\}$

yielding

(4) $E\{s^{-2}\} = E\{x_n^2\} - 2P^T H + H^T RH$

which is a quadratic function of H. By differentiating with respect to the elements of H we get

(5) $\nabla = -2P + 2RH.$

Setting $\nabla=0$ to find the optimal H , we get

(6) $H^* = R^{-1}P.$

The block solution optimal filter was calculated by solving equation 6. The filters were calculated using a standard Levinson's recursion algorithm [2].

Adaptive Solution

To calculate H^* adaptively a standard steepest descent algorithm is used:

(7) $H_{n+1} = H_n - \mu\nabla_n$

where the parameter μ controls convergence and stability. Since, we do not have access to V , we use a gradient estimate.

(8) $\hat{\nabla}_n = -2\bar{s}_n V_n$

which yields the algorithm

(9) $H_{n+1} = H_n + 2\mu\bar{s}_n V_n.$

By defining the expected value of H_n as M_n it can be shown that

(10) $M_n[I-2\mu R]^n H_0+R^{-1}P-[I-2\mu R]^n R^{-1}P.$

By diagonalizing R, it can be shown that

(11) $\lim_{n\to\infty}\{M_n\} = R^{-1}P$ for $0<\mu<\frac{1}{\lambda_{max}}$

where λ_{max} is the largest eigenvalue of the matrix R. The variance of the estimate can also be forced below any arbitrary positive limit as n gets large for V_k uncorrelated with V_j for $k \neq j$.

The optimal filter is a function of the inverse of R, $R^{-1}P$ (eq. 6). If R is singular it does not mean, in general, that there is no solution, simply that it is not unique. This condition is frequently encountered when the interfering noise is periodic, or nearly periodic. While channel estimation is not completely possible in such cases, it is only necessary to estimate the channel accurately in those frequency bands where significant interfering energy is present. Even though the channel estimate may be considered poor in such a situation, the noise reduction achievable may be significant.

Data Generation

If the data is generated as shown in figure 2, and if the channel is a finite length all-zero filter, perfect noise cancellation can be achieved if the estimated linear filter , H , converges to G . A more realistic model for data generation is given in Figure 3. The noise cancellation problem is then reduced to estimating of $G_2{-1}G_1$, see Figure 4. If G_1 and G_2 can be modelled as all- zero filters, the difficulty in estimating the optimal filter arises because of the need to effectively invert G_2. In general G_2 will not be a "minimum phase" process. Its inverse will, therefore, have poles outside the unit circle. For the estimated optimal filter to be stable will require it to be noncausal and doubly infinite. If its poles are well away from the unit circle, the response will be dominated by rapidly decaying exponentials. This allows us to approximate the required doubly infinite recursively generated filter, with a finite transversal filter. As the zeroes of G_2 and the actual poles of G_2^{-1} approach the unit circle, however, the number of points which we must allow in the active interval of the filter to be estimated grows if we desire to maintain a constant error, [3].

Basic Experiments

A white noise generator was used as a primary noise source. Its output was low-pass filtered to 3.2 KHz and sampled at a rate of 6.67 KHz. A square wave generator was used to generate nearly periodic noise sample. This sample was made highly non-stationary by varying the frequency adjustment of the square-wave generator in a semi-random fashion while digitizing. These samples were then concatenated and used as noise sources for both synthetic and acoustically recorded experiments.

Four FIR channel filters were used in order to analyze the performance of the noise cancellation. A low-pass filter with its cutoff frequency at approximately 1500 Hz and a triple band-pass filter were created. Two room-channel estimates were made from actual measurements of a room's response in order to simulate, digitally, an actual room [4].

Synthetic Experiments

For the initial experiments digitally recorded speech signals were added to the channel filtered noise segments and used as the noisy signals applied to the ANC algorithm. The corresponding unfiltered noise segments were used as the noise reference input signals. When white Gaussian noise was used as the interfering noise, accurate channel estimates were obtained,(H converged to G_2) for both low-pass and multi-band-pass channels. When the highly correlated, nearly periodic noise samples were used, the channel estimates did not converge to the known channels, but essentially complete noise cancellation still occurred.

Room Simulations

Using the measured room impulse responses, the degree of cancellation possible in a hard-walled room about fifteen feet square was determined. In the first experiment, the original noise signal was used as the reference noise(G_2=I), and one of the room channel filtered signals was used as the noisy signal. While the original room channel estimates were 4096 points long, the adaptive filter was constrained to a length of 3000 points. An adaption time-constant of approximately 0.4 seconds was specified. Noise reduction of -25 dB was measured for this experiment.

In the second experiment, the reference noise was generated by convolving the white noise with one of the room transfer functions,G_2 , while the noise added to the speech was generated by convolving the white noise through the other room transfer function,G_1 . This data model corresponds to Figure 3. Again 3000 points were specified for the adaptive filter's length, half of them before t=0. The resulting noise reduction measured was -12 dB.

Acoustically Recorded Experiments

Two similar experiments were performed in an actual acoustic environment. The digitized noise sources were played through a single multi-element

BOSE loudspeaker and digitally recorded through two separate SONY ECM-270 microphones placed at the same locations in the hard-walled room.

First a single channel room estimation experiment was performed . The reference noise was picked up by directly digitizing the speaker signal(G_2=I). The acoustically added speech plus noise signal was simultaneously recorded. The noise reduction achieved in this experiment was -24 dB.

Many experiments were performed where both signals were acoustically recorded. Using this data the optimal filter was estimated using both block and adaptive analysis. The results of these experiments for various filter lengths are compared in Table 1.

Observations on Results

Examination of the adaptive filter's impulse responses for the synthetic experiments showed that their estimates of the channels were excellent in frequency bands where significant noise energy was present, and very poor where no noise was present. This was not unexpected, since the adaptive filter's impulse response is a linear combination of previous reference noise sample vectors. For periodic noise the optimal filter was not unique , however the noise reduction was as good as that achieved when a white noise source was employed.

A comparison of the results obtained from synthetic and actual rooms (-25 dB vs. -24 dB in the single channel case, and -12 dB vs. -10.5 dB for the two channel case) indicates that the assumption that a room can be modelled as a linear, stationary channel appears valid. The results also show that spatially distributed noise sources, such as multi-element loudspeakers, do not cause of great deal degradation in performance. The comparisons of filter length verses noise reduction show relative performance losses caused by filter truncation. They are applicable to a single, hard-walled room about fifteen feet square, and ought not to be considered universally attainable levels. The absolute noise reduction obtainable for a given filter length is extremely dependent upon the physical environment where the process is being employed.

The comparisons of the ANC approach and the global block analysis showed that the adaptive procedure consistently performed better due to the nonstationarity of the noise. The block analysis was not developed for nonstationary data and attempted to minimize the total output energy. Also Levinson's recursion blew up when trying to compute a 3000 point filter.

ACKNOWLEDGEMENT

The author would like to thank Eleen Collins for preparation of the manuscript. This research was supported by the Information Processing Branch of the Defense Advanced Research Projects Agency, monitored by the Naval Research Laboratory under contract No. N00173-C-0041.

REFERENCES

[1] B.J. Widrow, et al., "Adaptive Noise Cancelling: Principles and Applications," Proceedings of the IEEE, vol. 63, pp. 1692-1719, Dec 1975.

[2] N. Levinson, "The Wiener RMS Error Criterion in Filter Design and Prediction", Journal of Mathematics and Physics, vol. 25, No.4,pp. 261-278, 1947.

[3] D. Pulsipher, "Application of Adaptive Noise Cancellation to Noise Reduction in Audio Signals," Ph.D. dissertation, University of Utah, 1978.

[4] S.F. Boll, E. Ferretti, T. Petersen, "Improving Synthetic Speech Quality Using Binaural Reverberation", Conference Report of *the 1976 IEEE ICASSP, pp,* . 705-708, 1976.

Filter Length	Block	Adaptive
10	0	-2
100	-3	-3.5
400	-4	-4.5
700	-5	-6.0
1500	-6.5	-8.0
3000	-3	-10.5

TABLE 1

Filter Length vs. Noise Reduction in dB.

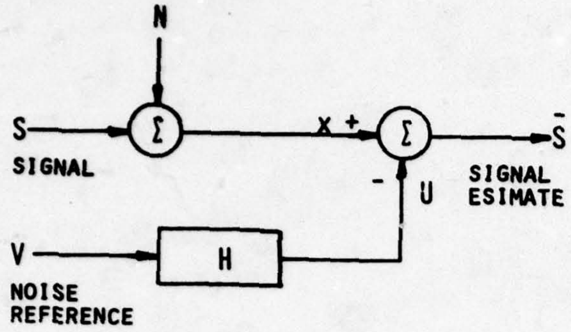


Figure 1 NOISE CANCELLING MODEL

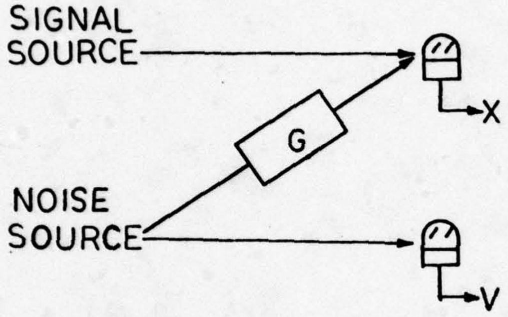


Figure 2 BASIC DATA GENERATION MODEL

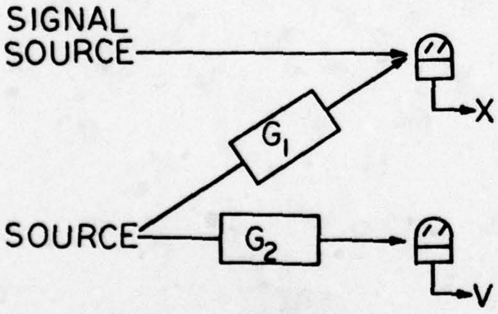


Figure 3 REALISTIC DATA GENERATION MODEL

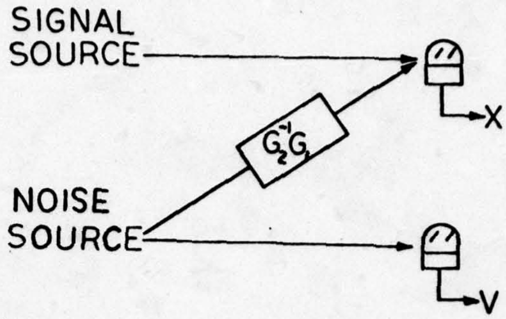


Figure 4 EQUIVALENT DATA GENERATION MODEL

RANK-ORDER SPEECH CLASSIFICATION ALGORITHM
(RASCAL)

Benjamin V. Cox

LaMar K. Timothy

December 1978

To be presented at ICASSP-79
April 2-4, 1979
Washington, D.C.

RANK-ORDER SPEECH CLASSIFICATION ALGORITHM*

(RASCAL)

Benjamin V. Cox

Sperry Univac, ASD, Salt Lake City, Utah

LaMar K. Timothy

University of Utah and Sperry Univac, ASD, Salt Lake City, Utah

ABSTRACT

This paper describes a theoretical and experimental investigation for detecting the presence of speech in wideband noise. A robust algorithm for making the silence-voiced-unvoiced decision is described. This algorithm is based on a nonparametric rank-order statistical signal-detection scheme that does not require a training set of data and maintains a constant false-alarm rate for a broad class of noise inputs corresponding to a single decision threshold. The nonparametric rank-order decision procedure is the multiple use of the two-sample Savage T statistic. The performance of this detector is evaluated and compared to that obtained by manually classifying twenty recorded utterances with 39, 30, 20, 10, and 0 decibel signal-to-noise ratios. In limited testing, the average probability of misclassification is less than 5 percent, 12 percent, and 55 percent for signal-to-noise ratios of 39, 20, and 0 decibels respectively.

INTRODUCTION

The fundamental problem in many speech communication and understanding systems is the search for a decision procedure that will classify speech in a noisy environment as voiced, unvoiced, or silence (noise alone). For several years, the notable advances in narrowband vocoders have motivated investigation into the theoretical aspects of robust speech classification algorithms that will effectively operate in adverse noise environments.

A number of papers and reports have been published describing the theory and techniques for making the voiced-unvoiced-silence (V-UV-S) classification [1]; however, very few papers have dealt with the problem of developing effective algorithms for real noise environments. In most of these papers, the detection of speech in background noise was conducted in a realtively noise-free environment under ideal laboratory acoustic recording conditions. The decision procedure that has enjoyed the widest acceptance is the pattern recognition approach of Atal and Rabiner [2]. This technique has been modified by various investigations [3]. The pattern recognition approach to the V-UV-S classification has usefulness for many speech processing system applications. However, it does not address the robustness issue in a communication sense since the technique requires a training set of data and will operate without degradation in performance only for a particular communication channel.

McAulay [4] has suggested an algorithm for detecting speech in an airborne command post noise environment, but it requires a large amount of signal processing, a speech-free interval to determine noise detection thresholds, and has not as yet been extensively tested.

The V-UV-S decision is a difficult problem in real noise environments; there is a need for continued research on the theory, techniques, and devices in this area [5].

In the research described here, a nonparametric rank-order statistical decision procedure that is theoretically recognized as robust in a communication sense has been formulated and investigated with a manually classified speech data base. It is theoretically robust in the communication sense since it has the desirable property of maintaining a constant false-alarm rate (CFAR) for a wide variety of noise distributions. The decision threshold is set independent of signal-to-noise ratio [6].

The detection performance for the Savage two-sample nonparametric rank-order test for speech signals in wideband noise is presented in this paper. A simple version of the problem is chosen in order to make a rigorous analysis possible, to evaluate the applicability of nonparametric procedures to V-UV-S classification, and to gain clarity.

Although this detection approach is new to speech processing, it is a mature statistical discipline. The nonparametric detection review paper by Thomas [7] indicates that a bibliography published in 1962 gives more than 3000 references. The application and analysis of nonparametric detections historically has been confined to nonengineering problems; an engineering text has only recently been published [8].

Some specific advantages of nonparametric statistics applied to speech detection are: (1) It maintains a constant false-alarm rate with a fixed threshold for large classes of noise distributions. (2) It does not require statistical information about either the signal or the background noise (does not require a training set of data) to set a decision threshold. (3) Performance for signals in non-Gaussian noise may often surpass that of detection optimized against Gaussian noise. (4) It will operate where the noise statistics are nonstationary or change from one application to another. (5) It can be digitally implemented.

DESCRIPTION OF THE ALGORTIHM

System Description

The system operates in the following manner: The speech signal is low-pass filtered to 3.2 kHz (telephone bandwidth), sampled at a 6.67-kHz rate, and high-pass filtered at approximately 200 Hz to remove any dc or low-frequency hum. The output from the high-pass filter is formulated into blocks of 100 samples (15 milliseconds of speech data). Each block of speech is then applied to four subband digital filters. The time

*Partially supported by DARPA, Information Processing Branch, contract N00173-77-C-0041 and monitored by NRL.

series output of each filter is labeled, pooled, and rank ordered. The rank-order values are then passed to the detector or classifier algorithm. Figure 1 shows a block diagram of the detection algorithm.

The filter subband partitioning is based on the work of Crochiene [9]. The important property achieved by this filter bank is that the sum of the individual frequency responses of the bandpass filters (composite response) lies flat with linear phase. The design of the subbands is based on perceptual criteria. The band-partitioning is such that each subband contributes equally to the articulation index (AI). The AI indicates, on the average, the contribution of each part of the spectrum to the overall perception of the spoken sound. Figure 2 shows the partitioning of the speech spectrum into four contiguous bands. These filters were designed using McClellan, Parks, and Rabiner's program [10].

Detection Procedure

To evaluate the applicability of nonparametric rank order detectors to the V-UV-S classification problem; three assumptions were made: (1) The spectrum of speech is different from bandlimited white noise. (2) The noise spectrum is approximately flat. (3) The amplitude distribution of speech is approximately Laplacian [11]:

$$p(x) = \frac{1}{\sqrt{2}\,\delta} \exp\left(-\sqrt{2}\frac{|x|}{\delta}\right) \qquad (1)$$

where δ is the rms speech value.

The detector based upon these assumptions operates in the following manner: The noise spectrum is assumed to be approximately flat over the telephone band of 200 to 3200 Hz. This frequency band is analyzed by forming four contiguous subbands. The subbands are chosen so that each subband data block is independent. A two-sample test statistic is used for each subband data block. The time-sampled data in the subband being tested forms the first sample, and the remaining pooled data forms the second sample. The procedure for the two-sample problem is to combine or pool both samples into a single ordered sample and then assign ranks [1, 2, . . . , N] to the sample values from the smallest to the largest value, without regard to the subband source of each value. The simplest test statistic is the sum of ranks assigned to the values from one of the subbands. If the sum (test statistic) is too large, there is some indication that the values from that subband tend to be larger than the values of the pooled second sample. The null hypothesis H_0 of no difference between subbands may be rejected if the ranks associated with one sample tend to be larger than those of the other sample; and the alternate hypothesis H_1 is accepted. Under the assumption that the rank of any single outcome is equally likely, the probability of any test statistic can be determined by counting outcomes, knowing there are N! total permutations. A test statistic for each subband data block is calculated and compared with a threshold determined from statistical tables. This decision procedure is referenced in mathematical literature as simultaneous statistical inference and is described more fully in [12]. The V-UV-S decision is a single-sided hypothesis test using the upper tail of the distribution function.

The detector compares a set of m time data samples from one of the subbands with pooled data from the other subbands to determine if the sample amplitude distributions (AD) are the same or different based on ranks. The form and parameters of the ADs are unknown. If the sample amplitude distributions in the subbands are statistically similar, to within testing error, noise only is declared at the output of the detector. If the sample AD in a subband with a frequency range below 2000 Hz is statistically different from the subbands forming the pooled sample, then voiced speech is declared at the output of the detector. If the opposite condition exists, then unvoiced speech is declared at the output of the detector. The decision procedure tested is closely related to the nonparametric detection procedure using a spectral data concept first introduced by Woinsky [13].

Test Statistic

The following description of the Savage test statistic follows the development presented in Hajek [14]. Since the amplitude distribution of speech is nearly exponential, the Savage test statistic is selected because it is the optimum rank statistic for an exponential distribution and a scale alternative. The Savage test statistic has the form

$$S = \sum_{i=1}^{N} A_i Z_i \qquad (2)$$

where Z_i is a switching function:

$$Z_i = \begin{matrix} 1 \text{ if the ith rank belongs to the filter output under test} \\ 0 \text{ otherwise} \end{matrix}$$

and where

$$A_i = \sum_{j=N-i+1}^{N} \frac{1}{j} \qquad (3)$$

which heavily weighs the ranks near the upper tail in the critical decision region. Under H_0 the Savage statistic satisfies

$$E(S) = m \,, \quad N = m + n$$
$$\mathrm{Var}(S) = \frac{mn}{N-1} \quad \left(1 - \frac{1}{N}\sum_{j-1}^{N}\frac{1}{j}\right) \qquad (4)$$

Consider the amplitude distribution function F (·) with standard deviation δ and zero mean corresponding to H_0. For the condition $\delta_1 > \delta_0$ we have $F(x/\delta_1) \leq F(x/\delta_0)$ in the critical test region of the upper tail Let δ_1 correspond to the sample from the subband under test and let δ_0 correspond to the pooled population under H_0. If voiced or unvoiced speech is present, then $\delta_1 > \delta_0$, otherwise $\delta_1 \leq \delta_0$. Consequently the hypothesis test can be stated as

$$\begin{aligned} H_0&: \ \delta_1 \leq \delta_0 \\ H_1&: \ \delta_1 > \delta_0 \end{aligned} \qquad (5)$$

which can be tested with ranks based on the Savage statistic S and a threshold S^{α} selected from rank statistics. The procedure follows.

Under the null hypothesis H_0, select a threshold S^{α} such that $P(S \leq S^{\alpha}) = 1 - \alpha$ where α is the probability of a type I error usually between 0.10 and 0.01. The quantiles of S^{α} are given in Table X of Hajek [14] for $N \leq 20$. If $N > 20$ a normal distribution approximation can be used considering Eq. 1 and 2. Accept H_0 if $S \leq S^{\alpha}$. Otherwise reject H_0.

Ranking of the data and calculation of S may require an excessive number of computer manipulations; the procedure requires that all data from the filters be stored for each data frame for ranking purposes. This problem can be reduced using a mixed Savage statistical test [14] which was applied to the data presented in the following section.

EXPERIMENTAL RESULTS

The nonparametric classifier was tested on the diagnostic rhyme test (DRT) file tape supplied by Dyna Stat Incorporated [4]. The additive white noise tape was generated by digitizing the analog output of an analog noise generator. Both the word file and the noise file were prefiltered with a low-pass filter having a 3.2 kHz cutoff frequency and were sampled at 6.667 kHz.

Using the software programs described in [15] and the DRT data file, a controlled DRT word data base with additive white noise of progressively smaller signal-to-noise ratios: 39, 30, 20, 10, and 0 dB were created and processed by the detector algorithms. Tests were conducted to evaluate the speech detector's performance for the five different signal-to-noise ratios of wideband Gaussian noise. For each clean test word from the DRT file, a manual analysis was performed on each 15-millisecond interval to classify it as voiced, unvoiced, or silence based on visual inspection of the acoustic waveform and a phonetic interpretation of the utterance. Two independent manual classifications were made on each test word.

A V-UV-S decision was made by the computer every 15 milliseconds based on a mixed Savage statistic using 100 samples from each filter subband represented in figure 2. The mixed Savage statistic was formed by averaging the absolute value of 5 samples forming 20 averaged samples per subband. The 80 averages from the four subbands were pooled and ranked.

Error rates were computed by comparing the manual classification with the detector's classification output. Table 1 summarizes the overall recognition rate as a function of signal-to-noise for the simultaneous decision procedure for all 20 test utterances.

The recognition results in Table 1 and spectral analysis of the DRT background noise indicate that a significant low-frequency spectral component is present in the background noise of the DRT file. Table 1 and the additional test described in [1] show that as noise is added, the effect is to whiten the spectrum, and therefore, the misclassification rate decreases at 30 dB as compared to 39 dB.

CONCLUSIONS

A theoretical and experimental investigation for detecting the presence of speech in wideband noise and classifying the detected utterance as voiced or unvoiced, based on a nonparametric statistical detection approach, has been described. The speech detection technique that was tested is effective for detecting speech in wideband noise at a signal-to-noise ratio from 39 to 0 dB and meets the requirement for being independent of transmission channel characteristics, recording conditions, and distribution of the background random noise. The desirable feature of this detection or classification scheme is that is requires neither a training set of data nor a priori information of the statistical parameters of speech or background noise.

ACKNOWLEDGMENTS

Gratitude is expressed to Dr. S. F. Boll and Dr. C. K. Rushforth at the University of Utah for support and helpful suggestions during the course of this investigation.

REFERENCES

(1) B V Cox, "The Application of Nonparametric Rank-Order Statistics to Robust Speech Activity Detection", Ph.D. dissertation, Dep. EEC. ENG., Univ. of Utah, Dec. 1978.

(2) Bishnu S. Atal and L. R Rabiner, "A Pattern Recognition Approach to Voice-Unvoiced-Silence Classification with Application to Speech Recognition", IEEE TRANS on Acoustics, Speech, and Signal Processing, Vol. ASSP-24, No. 3, June 1976, pp. 201-212.

(3) V. V. S. Sarma and D. Vengopal, "Studies on Pattern Recognition Approval to Voiced-Unvoiced-Silence Classification", Conference Record, 1978 IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE Cat. No. 78CH1285-6 ASSP, Tulsa, Ok., April 1978, pp. 1-4.

(4) R. J. McAulay, "Optimum Classification of Voice Speech, Unvoiced Speech, and Silence in the Presence of Noise and Interference", Technical Note 1976-7, Lincoln Laboratory, MIT, 3 June 1976.

(5) B. Gold, "Digital Speech Networks", Proceedings of the IEEE, Vol. 65, No. 12, December 1977, pp. 1636-1658.

(6) R. F. Daly and C. K. Rushforth, "Nonparametric Detection of a Signal of Known Form in Additive Noise", IEEE TRANS on Information Theory, Vol. IT-11, Jan. 1965, pp. 70-76.

(7) J. B. Thomas, "Nonparametric Detection", Proceeding of the IBEE, Vol. 58, No. 5, May 1970, pp. 623-631.

(8) J. D. Gibson and J. L. Melsa, Introduction to Nonparametric Detection with Applications, New York: Academic Press, 1975.

(9) R. E. Crochiere and M. R. Sambur, "A Variable Band Coding Scheme for Speech Encoding at 4.8 kb/s", *Conference Record, 1977 IEEE International Conference* on Acoustics, Speech, and Signal Processing, IEEE Cat. No. 77CH1197-3, Hartford, Conn., May 1977, pp. 444-447.

(10) J. H. McClellan, T. W. Parks, and L. R. Rabiner, "A Computer Program for Designing Optical EIR Linear Phase Digital Filters", IEEE TRANS on Audio and Electro Acoustics, Vol. AG-21, December 1973, pp. 506-526.

(11) M. D. Paez and T. H. Glisson, "Minimum Mean-Squared-Error Quantization in Speech PCM and DPCM Systems", IEEE TRANS on Communications, Vol. COM-20, April 1972, pp. 225-230.

(12) J. D. Gibbons, Nonparametric Statistical Inference, New York: McGraw-Hill, 1971.

(13) M. W. Woinsky, "Nonparametric Detection Using Special Data", IEEE TRANS on Information Theory, Vol. IT-18, No. 1, January 1973, pp. 110-118.

(14) Jaroslav Hajek, A Course in Nonparametric Statistics, San Francisco: Holden-Day, 1969.

(15) S. F. Boll, et al, "Noise Suppression Methods for Robust Speech Processing", Semi-Annual Technical Report. UTEC-CSC: 77-202, Computer Science Dept., University of Utah, April 1977.

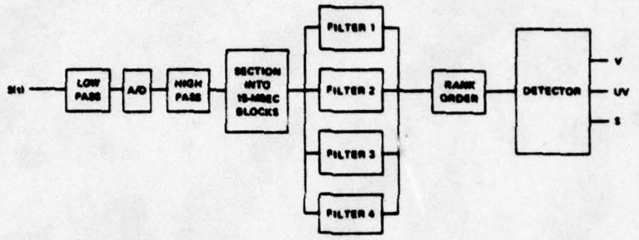


Figure 1. Block Diagram of Signal Classification Method

SUB-BAND NUMBER	FREQUENCY RANGE (Hz)
1	200 – 700
2	700 – 1310
3	1310 – 2030
4	2030 – 3200

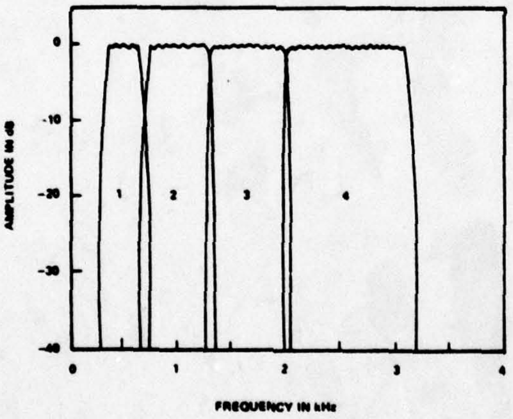


Figure 2. Partitioning of the Speech Spectrum into Four Contiguous Bands that Contribute Equally to Articulation Index with Frequency Ranges of 200 to 3200 Hz.

Table 1.

20-Sample Recognition Rate for the Simultaneous Decision Procedure

Percent Recognition	Silence					Voiced					Unvoiced				
S/N in dB	39	30	20	10	0	39	30	20	10	0	39	30	20	10	0
Gob	-	-	-	-	-	97	95	79	51	28	-	-	-	-	-
Sue	61	-	100	100	100	100	-	100	100	100	100	-	92	58	25
Taunt	82	95	95	95	91	95	95	95	82	36	100	100	100	0	0
Nil	23	85	100	100	100	97	100	89	78	49	-	-	-	-	-
Boast	46	82	96	89	89	100	100	95	84	58	100	100	67	0	0
Jab	70	90	90	90	90	84	84	70	38	24	75	100	75	50	25
Cheat	68	91	95	91	91	88	100	91	91	76	100	86	86	71	57
Said	57	71	86	100	100	100	100	93	52	44	-	-	-	-	-
Gnaw	0	75	100	100	100	100	100	94	86	17	-	-	-	-	-
Weed	100	100	100	100	100	98	95	93	79	45	-	-	-	-	-
Deck	78	100	100	100	100	95	82	77	59	41	86	43	29	0	0
Chew	100	100	100	100	100	93	96	90	90	45	86	86	86	71	43
Thong	100	100	100	100	100	95	95	86	84	22	-	-	-	-	-
Keep	89	100	100	100	100	94	94	88	71	71	100	100	67	33	0
Got	86	90	95	86	86	91	83	70	61	30	100	100	100	0	0
Dank	91	100	100	100	100	92	89	78	50	28	-	-	-	-	-
Shoes	86	100	-	100	100	100	100	-	100	77	100	100	-	83	50
Shag	33	67	100	100	100	94	97	87	61	42	100	100	91	64	27
Pool	63	88	100	100	100	97	97	95	86	51	-	-	-	-	-
Dip	59	91	95	100	100	96	87	83	48	26	-	-	-	-	-
Average Percent Recognition	68	90	97	97	97	95	94	87	72	45	95	91	79	47	20

ESTIMATING THE PARAMETERS OF A NOISY ALL-POLE PROCESS USING POLE-ZERO MODELING

W. J. Done

C. K. Rushforth

December 1978

To be presented at ICASSP-79
April 2-4, 1979
Washington, D.C.

ESTIMATING THE PARAMETERS OF A NOISY ALL-POLE PROCESS USING POLE-ZERO MODELING

W. J. Done
Amoco Research Center
Tulsa, Oklahoma 74102

C. K. Rushforth
Department of Electrical Engineering
University of Utah
Salt Lake City, Utah 84112

ABSTRACT

Linear predictive coding (LPC) has been successfully applied to the encoding of speech and other time series. It has been widely observed, however, that the performance of an LPC algorithm deteriorates rapidly in the presence of background noise. In this paper, we describe and discuss one approach to the identification of a time series corrupted by additive white noise.

A common approach to this problem is to prefilter the noisy time series, and then to apply an estimation algorithm which treats the time series as if it were noise-free. We describe an alternative approach which involves modifying the time-series model at the outset to account for the presence of noise. An estimation algorithm is then developed for this modified model. We discuss the development of the model, the estimation algorithm, and some representative experimental results.

INTRODUCTION

Linear Predictive Coding (LPC) has been widely and successfully applied to the encoding and processing of speech waveforms and other time series. Most of the initial demonstrations of LPC were conducted using high-quality and relatively noise-free signals, however. It has recently become clear that background noise and other perturbations can cause a serious degradation in the performance of LPC algorithms (1, 2). In speech processing, for example, the presence of noise can adversely affect silence detection, voiced/unvoiced determination, pitch period calculation, and identification of the LP coefficients. The work discussed in this paper deals only with the problem of coefficient identification, and is applicable to any time series which can be modeled as an all-pole or autoregressive (AR) process perturbed by additive white noise. We make the further simplifying assumption that the order of the process is known; thus, only the unknown coefficients of the difference equation defining the AR process must be estimated from the observed data.

Several schemes have been developed to deal with the effects of noise on LPC estimation algorithms (1, 3, 4). The approach to noisy time-series analysis which we discuss in this paper involves a modification of the process model at the outset to account for the effects of additive white noise. We show that the addition of white noise to an AR(q) process (an all-pole process with q poles) results in a new process which is an autoregressive moving-average (ARMA) process with q poles and q zeroes. Furthermore, the poles of the new ARMA (q, q) process are identical to the poles of the original AR(q) process, a fact which greatly simplifies the estimation process. By modifying the model in this way, we transform the problem of estimating the parameters of an AR process in the presence of noise into a problem of estimating the parameters of an ARMA process which has the same poles or AR coefficients as the original process.

Optimal estimation of the parameters of an ARMA process is much more difficult than estimating the parameters of an AR process. Our objective in this paper is to determine whether there is any performance advantage to be gained using the approach described above, and we do not concern ourselves with computational efficiency per se. If this method were to be implemented, it would no doubt have to be modified to increase its speed.

Estimation of the parameters of an ARMA process has been extensively discussed in the literature (5, 6, 7). We describe an algorithm developed by Anderson (7) for conditional maximum-likelihood estimation using a version of the Newton-Raphson method.

Finally, we present the results of a number of experiments conducted using simulated time-series data. We compare the estimates obtained using the Newton-Raphson method with those obtained by applying the standard autocorrelation method of LPC estimation to both the unmodified noisy time series and to a Wiener-filtered version of this noisy time series. We also include results obtained by solving the "shifted" Yule-Walker equations (8).

THE MODEL

In this section, we give a very brief development of the model which results when white noise is added to an AR process. For more details, see (9) or (10).

We assume that the desired signal process s(k) is a stationary AR(q) process of known order q described by

$$\sum_{i=0}^{q} a(i)s(k - i) = \varepsilon(k), \tag{1}$$

where $\varepsilon(k)$ is a sequence of independent, identically-distributed Gaussian random variables with mean zero and variance σ_ε^2. We further assume that a(0) = 1 and that q > 0. This signal process is contaminated by a sequence n(k) of independent, identically-distributed Gaussian random variables with mean zero and variance σ_n^2 to form the observable sequence

$$x(k) = s(k) + n(k). \tag{2}$$

It can be shown (8, 9) that x(k) satisfies the relationship

$$\sum_{i=0}^{q} a(i)x(k - i) = \sum_{j=0}^{q} b(j)v(k - j), \tag{3}$$

where b(0) = 1 and v(k) is a sequence of independent, identically-distributed Gaussian random variables with mean zero and variance σ_v^2. Thus, the observed noisy process x(k) can be viewed as an ARMA (q, q) process with AR coefficients $\{a(i)\}_{i=1}^{q}$, MA coefficients $\{b(j)\}_{j=1}^{q}$, and driving-sequence variance σ_v^2. This new model contains 2q + 1 parameters compared with q + 2 for the original model.

Upon comparing (1) and (3), we see that the AR coefficients of the new ARMA model are identical to those of the desired signal process s(k). Hence, after estimating the parameters of the ARMA process, we can simply discard the MA estimates and retain the AR estimates. This result rests on the assumption that the additive noise is white. If it is not, a similar result can be established but the AR parameters will no longer be the same.

PARAMETER ESTIMATION

We showed in the previous section that estimation of the parameters of an AR process contaminated by additive white noise can be accomplished by estimating the parameters of an associated ARMA process. We have adopted and implemented an ARMA estimation algorithm of Anderson (7) for this purpose, and we briefly describe this algorithm in this section.

Of the several methods described in (7), the one we selected is the so-called *time-domain Newton-Raphson method*. To begin, we define the N × N matrix

$$\underline{L}^k = \begin{bmatrix} 0 & 0 \\ I_{N-k} & 0 \end{bmatrix} \tag{4}$$

where I_{N-k} is the (N - k) × (N - k) identity matrix. Further, define column vectors $\underline{x} = [x(0)\ldots x(N - 1)]^T$ and $\underline{v} = [v(0)\ldots v(N - 1)]^T$. Then $\underline{L}^k\underline{x} = [0\ldots0\ x(0)\ldots x(N - 1 - k)]^T$.

Using the matrices $\underline{L}^k$, and assuming that x(k) = v(k) = 0 for k < 0, we can write (3) in matrix form

$$\underline{A}\ \underline{x} = \underline{B}\ \underline{v} \tag{5}$$

where

$$\underline{A} = \sum_{i=0}^{q} a(i)\ \underline{L}^i \tag{6}$$

and

$$\underline{B} = \sum_{j=0}^{q} b(j)\ \underline{L}^{(j)}. \tag{7}$$

The conditional log likelihood function (conditioned upon the inital values assumed for x(k)) can now be written

$$\ell n(f) = -\frac{N}{2}\ell n(2\pi) - \frac{N}{2}\ell n\left(\sigma_v^2\right) + \ell n\ |\underline{A}| - \ell n\ |\underline{B}| - \frac{1}{2\sigma_v^2}\underline{x}^T\ \underline{A}^T\left(\underline{B}^T\right)^{-1}\ \underline{B}^{-1}\ \underline{A}\ \underline{x}. \tag{8}$$

The conditional maximum likelihood estimates for $\{a(i)\}_{i=1}^{q}$, $\{b(j)\}_{j=1}^{q}$, and σ_v^2 can be obtained in principle by differentiating (8) with respect to each of these parameters, equating the results to zero, and solving the resulting set of simultaneous equations. Unfortunately, these equations are nonlinear in the parameters and cannot be solved directly. Thus, we must resort to iterative methods of solution.

The estimation of σ_v^2 can be decoupled from the estimation of the a(i) and b(j). Specifically, if we use (5) to define

$$\underline{v} = \underline{B}^{-1}\ \underline{A}\ \underline{x}, \tag{9}$$

then

$$\hat{\sigma}_v^2 = \frac{1}{N}\underline{v}^T\ \underline{v}. \tag{10}$$

To estimate the a(i) and b(j), we define $\underline{a} = [a(1)\ \ldots\ a(q)]^T$, $b = [\underline{b}(1)\ \ldots\ b(q)]^T$, and $\underline{\theta} = [a^T \vdots b^T]^T$. The estimate of $\underline{\theta}$ is obtained iteratively using the equation

$$\underline{\theta}_{i+1} = \underline{\theta}_i + \underline{R}_i^{-1}\ \underline{g}_i, \tag{11}$$

where $\underline{g}_i$ is the gradient vector and $\underline{R}_i$ is a matrix whose elements will be given below. To use (11), an initial value $\underline{\theta}_o$ is chosen, $\underline{R}_o$ and $\underline{g}_o$ are calculated, and these values are then used to obtain an updated estimate $\underline{\theta}_1$. The process is then repeated iteratively until some stopping criterion is satisfied.

It is convenient to express $\underline{R}$ and $\underline{g}$ in the partitioned forms

$$\underline{R} = \begin{bmatrix} \underline{\Phi} & \underline{\Omega} \\ \underline{\Omega}^T & \underline{\Psi} \end{bmatrix} \tag{12}$$

and

$$\underline{g} = \begin{bmatrix} \underline{w} \\ \underline{u} \end{bmatrix}. \tag{13}$$

The vectors $\underline{w}$ and $\underline{u}$ are q × 1 column vectors whose jth and mth elements are, respectively,

$$[w]_j = \frac{1}{\sigma_v^2} \underline{v}^T \underline{L}^j \underline{B}^{-1} \underline{v} \tag{14}$$

and

$$[u]_m = - \frac{1}{\sigma_v^2} \underline{v}^T \underline{L}^m \underline{A}^{-1} \underline{v} \tag{15}$$

where $\underline{v}$ is computed from the observed data vector $\underline{x}$ by (9). The elements of $\underline{\Phi}$, $\underline{\Omega}$, and $\underline{\psi}$ are given by

$$[\Phi]_{jk} = \frac{1}{\sigma_v^2} \underline{v}^T \left(\underline{B}^T\right)^{-1} \left(\underline{L}^T\right)^j \underline{L}^k \underline{B}^{-1} \underline{v}, \tag{16}$$

$$[\Omega]_{jm} = \frac{-1}{\sigma_v^2} \underline{v}^T \left(\underline{B}^T\right)^{-1} \left(\underline{L}^T\right)^j \underline{L}^m \underline{A}^{-1} \underline{v}, \tag{17}$$

and

$$[\psi]_{mn} = \frac{1}{\sigma_v^2} \underline{v}^T \left[\underline{A}^T\right]^{-1} \left[\underline{L}^T\right]^m \underline{L}^n \underline{A}^{-1} \underline{v}. \tag{18}$$

To obtain $\underline{R_i}$ and $\underline{g_i}$, we simply substitute the parameters from the ith estimate $\underline{\theta_i}$ into (12)-(18). For a detailed discussion of the computations involved, see [8].

EXPERIMENTAL RESULTS

We have conducted extensive tests of the time-domain Newton-Raphson algorithm described in the previous section, and have compared its performance to that of several other estimators. In this section, we briefly summarize some representative results and discuss the conclusions we have drawn from these results. A discussion of results obtained using an algorithm developed by Steiglitz [11] appears elsewhere in these proceedings [12], and a much more extensive discussion of all our results appears in [8].

Although we have conducted tests on higher-order processes, we restrict our attention here to a first-order AR process contaminated by additive white noise. For definiteness, we took the single AR parameter to be 0.5. As an initial test of the performance of the Newton-Raphson algorithm we performed, for a number of 256-point frames of data, a straightforward search in (a, b) parameter space to locate that point which minimized the unconditional sum of squared residuals (see (10), Chapter 7). The Newton-Raphson algorithm was applied to the same data, and its estimates of a and b were compared with the values obtained using the search procedure. In all cases in which the Newton-Raphson procedure converged, the results agreed very closely. These results confirm that when poor estimates are obtained using the NR method, it is almost always the case that these poor estimates really do minimize the sum of the squared residuals. Thus, the weakness is not in the NR algorithm per se, but is inherent in the underlying least-squares approach to estimation.

In most experiments using the NR algorithm, we used the true parameter values as the initial values. We did this on purpose in order to avoid extensive discussion of this issue. The primary effect of using other reasonable initial guesses should be a modest increase in the rate of failure to converge, and does not seriously affect our conclusions. This is borne out by some results obtained when we did not know the true values for the moving-average parameters and therefore were forced to use other starting values.

Estimates of the AR parameter a were obtained for 518 frames, each containing 256 points of data, for each of six signal-to-noise ratios. The methods used to obtain these estimates were the following:

1. The time-domain Newton-Raphson method described above.
2. The standard autocorrelation method of LPC.
3. Solution of the shifted Yule-Walker equations to account for the moving-average portion of the process.
4. Wiener filtering, assuming knowledge of the signal and noise spectra, followed by LPC estimation. In practice, these spectra are not known, and in fact are to be estimated, but this approach provides an indication of what can be achieved.

The results obtained using these four methods were averaged over the 518 frames of data, and these average estimates are plotted in Fig. 1. In the case of the NR method, the average was taken only over those frames for which convergence occurred (515 at 0 dB, 214 at -10 dB, 518 in all other cases). In terms of these average results, it is clear that the NR and shifted Yule-Walker methods are superior to the other methods. In particular, the SNR threshold below which the estimate becomes very poor is roughly 14 dB lower for the NR method than for the uncorrected LPC method.

Looking only at the averages can be somewhat misleading, however. A more complete picture is obtained by looking at the variances of the estimates, and here some of the advantage of the NR algorithm is lost. The variance of the NR estimate is appreciably larger than that of the LPC estimate, as is shown in Fig. 2. Thus, although the NR method is superior on the average, the LPC estimate will actually be better for a significant number of individual frames.

SUMMARY

In this paper, we have shown that an AR (all-pole) process contaminated by additive white noise can be modeled as an ARMA (pole-zero) process whose poles are identical to those of the original AR process. Thus, the problem of estimating the parameters of a noisy AR process can be transformed into one of estimating the parameters of an ARMA process, unfortunately a much harder problem.

We implemented an ARMA estimation algorithm using the Newton-Raphson approach, and then applied this algorithm to a large amount of data from a synthetic noisy AR (1) process. This procedure yielded considerably better results on the average than did an unmodified LPC algorithm, but this advantage is qualified by the fact that the variance of the NR estimate is greater than that of the LPC estimate.

ACKNOWLEDGMENT

This work was supported in part by the Information Processing Branch of the Defense Advanced Projects Research Agnecy, monitored by the Naval Research Laboratory, under contract No. N00173-C-0041.

REFERENCES

(1) B. Yegnanarayana, "Effect of Noise and Distortion in Speech on Parametric Excitation", *Proceedings of the 1976 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 336-339.

(2) M. R. Sambur and N. S. Jayant, "LPC Analysis/ Synthesis from Speech Inputs Containing Quantizing Noise or Additive White Noise", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 488-494:December 1976.

(3) S. F. Boll, "Improving Linear Prediction Analysis of Noisy Speech by Predictive Noise Cancellation", *Proceedings of the 1977 International Conference on Acoustics, Speech, and Signal Processing*, pp. 10-12.

(4) J. S. Lim and A. V. Oppenheim, "All-Pole Modeling of Degraded Speech", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 197-210:June 1978.

(5) E. J. Hannan, "The Estimation of Mixed Moving Average Autoregressive Systems", *Biometrika*, Vol. 56, No. 3, pp. 579-593:1969.

(6) H. Akaike, "Maximum Likelihood Identification of Gaussian Autoregressive Moving Average Models", *Biometrika*, Vol. 60, No. 2, pp. 255-265:1973.

(7) T. W. Anderson, "Estimation for Autoregressive Moving Average Models in the Time and Frequency Domains", *Annals of Statistics*, Vol. 5, No. 5, pp. 842-865:1977.

(8) W. J. Done, "Estimation of the Parameters of an Autoregressive Process in the Presence of Additive White Noise", Ph.D. Dissertation, Electrical Engineering Department, University of Utah, Salt Lake City, Utah, 1978.

(9) M. Pagano, "Estimation of Models of Autoregressive Signal Plus White Noise", *Annals of Statistics*, Vol. 2, No. 1, pp. 99-108:1974.

(10) G. E. Box and G. M. Jenkins, *Time Series Analysis, Forecasting, and Control*, Holden-Day, San Francisco, California, 1976.

(11) K. Steiglitz, "On the Simultaneous Estimation of Poles and Zeros in Speech Analysis", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 229-234:June 1977.

(12) W. J. Done and C. K. Rushforth, "Evaluation of the Steiglitz Algorithm for Estimating the Parameters of an ARMA Process", *Proceedings of the 1979 International Conference on Acoustics, Speech, and Signal Processing*.

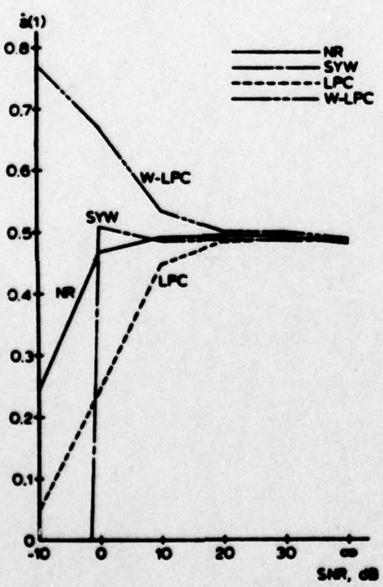


Fig. 1. Comparison of four estimators of a(1). True value = 0.5.

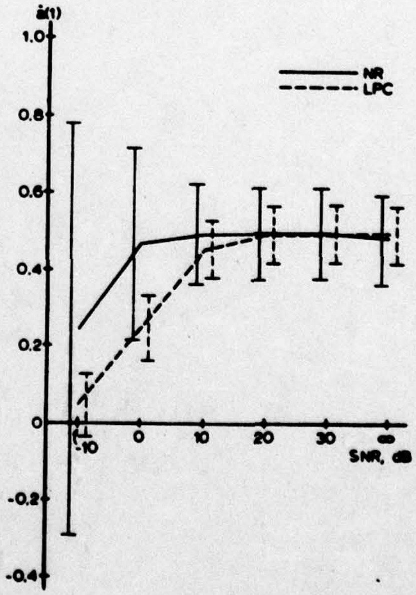


Fig. 2. NR and LPC estimates of a(1) with ±σ limits.

EVALUATION OF THE STEIGLITZ ALGORITHM FOR ESTIMATING THE PARAMETERS OF AN ARMA PROCESS

W. J. Done

C. K. Rushforth

December 1978

To be presented at ICASSP-79
April 2-4, 1979
Washington, D.C.

EVALUATION OF THE STEIGLITZ ALGORITHM FOR ESTIMATING THE PARAMETERS OF AN ARMA PROCESS

W. J. Done
Amoco Research Center
Tulsa, Oklahoma 74102

C. K. Rushforth
Department of Electrical Engineering
University of Utah
Salt Lake City, Utah 84112

ABSTRACT

Steiglitz (2) has recently described an algorithm for estimating the parameters of an autoregressive-moving-average (ARMA) process. This algorithm has application, for example, to the problem of determining the poles and zeroes of the vocal-tract transfer function.

In this paper, we report and discuss the results of a number of simulations conducted using the Steiglitz algorithm. The bulk of the experiments involved driving the ten-pole, two-zero filter described in (2) with a single pulse, with a short pulse train, and with samples of white Gaussian noise. In each of these cases, we evaluated the effects of such processing options as windowing, preemphasis, and cepstral-domain filtering. We also discuss and compare results obtained by applying the Steiglitz algorithm and a Newton-Raphson conditional maximum-likelihood algorithm to a first-order process.

INTRODUCTION

Steiglitz and McBride (1) propose a system identification procedure in which the z-domain transfer function of the unknown system is B(z)/A(z). B(z) and A(z) are polynomials given by

$$A(z) = \sum_{i=0}^{q} a(i)z^{-1}, \ a(0) = 1.0, \qquad (1)$$

and

$$B(z) = \sum_{i=0}^{p} b(i)z^{-1}, \ b(0) = 1.0. \qquad (2)$$

The polynomial A(z) determines the pole locations of the model and is, equivalently, the autofegressive (AR) operator. The zero locations are determined by B(z), the moving-average (MA) operator. Thus, if the driving sequence v(k) is a white noise sequence, the response x(k) is an ARMA process. Assuming that the input V(z) and output X(z) are known, the model's response is U(z) = [B(z)/A(z)]V(z). The error is then given by E(z) = U(z) - X(z). After linearizing the model, Steiglitz and McBride arrive at an iterative procedure which estimates the coefficients a(1), ..., a(q), b(0), ..., b(p).

In (2), Steiglitz applies this method to data obtained from a system in which the input V(z) is unknown. The data x(k) are assumed to reslut from driving the unknown system with an impulse. Steiglitz applies this method to a ten-pole, two-zero "unknown" system in which the input v(k) is actually an impulse train, simulating voiced speech. Because the data x(k) are assumed to result from an impulsive input, Steiglitz proposes that x(k) be modified prior to analysis to improve that assumption. Preemphasis, windowing, and cepstral-domain operations are suggested toward that end.

EXPERIMENTAL RESULTS

The application of this algorithm to data generated using white noise as the input to the ten-pole, two-zero model used by Steiglitz is reported in (3). This represents the situation usually encountered in ARMA model estimation. For comparison, the algorithm is also applied to data generated with inputs of a single impulse and an impulse train. The various modifications to x(k) proposed by Steiglitz are performed. The resulting sequence is analyzed to obtain estimates of the ARMA coefficients. Results are reported here on those modifications which produced estimates of the ARMA coefficients having the smallest mean square error when compared to the coefficients used to generate the data.

Figure 1 shows the spectrum (in dB) of the system to be identified. Using an input of a single impulse, the resulting data sequence x(k) is shown in Fig. 2. The best estimates of the ARMA coefficients are obtained by applying the algorithm directly to x(k). The spectrum of the estimated model after one iteration is shown in Fig. 3. There is essentially no error in the estimate.

The next case to be discussed is the analysis of data generated using an impulse train as the input. The impulses occur every 100 sample points. The resulting output is shown in Fig. 4. In this case, the following modifications are made to the data x(k):

1. Hamming window x(k).
2. Window the complex cepstrum of x(k).
3. Transform the resulting cepstral sequence to obtain the time domain sequence $x_{mp}(k)$.

After windowing x(k) in step 1 to obtain $w(k) \cdot x(k)$, the data are transformed using an N-point DFT. Prior to computing the complex cepstrum, the signal is forced to have zero-phase. The complex logarithm thus becomes

$$\mathrm{Re}\{\log X(\ell)\} = \frac{1}{2}\log\left\{[\mathrm{Re}\, X(\ell)]^2 + \mathrm{Im}[X(\ell)]^2\right\} \quad (3)$$

$$\mathrm{Im}\{\log X(\ell)\} = 0, \quad (4)$$

where $X(\ell)$ is the DFT of $w(k) \cdot x(k)$. After the windowing operations to be performed on the cepstrum, the zero-phase assumption yields the same results as those that would be obtained using the actual phase of $X(\ell)$. The imposition of zero phase avoids the necessity of phase unwrapping. The complex cepstrum $\hat{x}(k)$ of this zero-phase signal is found by performing another DFT on the complex logarithm.

Two operations are now performed on $\hat{x}(k)$. First, let $\hat{x}_{mp}(k) = u(k) \cdot \hat{x}(k)$, where

$$u(k) = \begin{cases} 1, & k = 0 \\ 2, & k = 1, \ldots, \frac{N}{2} \\ 0, & k = \frac{N}{2} + 1, \ldots, N \end{cases}$$

The cepstrum is now causal, and the corresponding time-domain signal has a magnitude spectrum identical to that of $w(k) \cdot x(k)$. The second operation performed on $\hat{x}(k)$ is to zero the portion of the cepstrum having the pitch spike associated with the periodic nature of x(k). The cepstral signal $\hat{x}_{mp}(k)$ resulting from these two procedures is transformed back to the time-domain signal $x_{mp}(k)$. The cepstral processing has achieved two goals:

1. $x_{mp}(k)$ is a minimum phase sequence.
2. The periodic nature of x(k) is suppressed.

The assumption of an impulsive input is more nearly valid for $x_{mp}(k)$ than for x(k). Analysis to determine the ARMA coefficients is performed on $x_{mp}(k)$, which is shown in Fig. 5.

It was found that the Hamming window step was necessary to obtain a convergent, stable filter estimate. Preemphasis was not performed on either x(k) or $x_{mp}(k)$, as this tended to degrade the coefficient estimates somewhat. The spectrum of the estimate after two iterations is shown in Fig. 6 and is quite good, confirming the results in (2).

The last case to be considered is for data generated when the input to the system is an approximately white noise sequence. The resulting output is shown in Fig. 7. The best estimates in this case were obtained by analyzing x(k) directly. Unlike the impulse excited case, however, the estimate is poor. Figure 8 shows the spectrum of the estimate after two iterations. Further iterations result in progressively narrower and higher spectral peaks. The estimate often becomes unstable. The algorithm is no longer achieving the excellent results found in the other two cases.

Because the tenth order AR operator is likely to tax any estimation algorithm, the algorithm developed by Steiglitz was applied to a single-pole system (q = 1, p = 0), excited by white noise. The single denominator coefficient, a(1), was 0.5 for this test. The estimate for a(1) from the Steiglitz algorithm is compared to that obtained from a Newton-Raphson (NR) implementation of a maximum likelihood ARMA estimation procedure (4). The results for ten iterations of one frame of data are presented in Table 1. The initial guess for a(1) in both algorithms is 0.5, the actual value of the coefficient used to generate the data. Using this as the initial guess removes the uncertainty about an initial guess from the test. From Table 1, we see that after the first iteration, the NR estimate does not change in at least the five most significant figures. The Steiglitz estimate, however, varies considerably from iteration to iteration. The estimate at iterations 2, 3, and 4 is unstable. Convergence does occur in later iterations, but to a value indicating the pole is close to the unit circle. This results in a narrow, high peak in the spectrum, characteristic of the estimate in the tenth order case. In addition, computations using the Steiglitz algorithm requiré the use of double precision arithmetic, even in the previous successful cases. The NR method does not require double precision arithmetic for successful parameter estimation.

CONCLUSION

The results of the tests performed here confirm that the parameter estimation algorithm proposed by Steiglitz (2) produces good results for the impulse- and impulse-train-excited cases. Care must be taken, however, in choosing modifications to x(k). The performance of the algorithm in the noise-excited case is poor, even for a first-order process. The algorithm does not appear to be applicable to the problem of estimating the parameters of a noise-excited ARMA process.

ACKNOWLEDGMENT

This work was supported in part by the Information Processing Branch of the Defense Advanced Projects Research Agency, monitored by the Naval Research Laboratory, under contract No. N00173-C-0041.

REFERENCES

(1) K. Steiglitz and L. E. McBride, "A Technique for the Identification of Linear Systems", *IEEE Transactions on Automatic Control*, Vol. AC-10, No. 4, pp. 461-464:July 1965.

(2) K. Steiglitz, "On the Simultaneous Estimation of Poles and Zeros in Speech Analysis", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-25, No. 3, pp. 229-234: June 1977.

(3) W. J. Done, *Estimation of the Parameters of an Autoregressive Process in the Presence of Additive White Noise*, Ph.D. Dissertation, Department of Electrical Engineering, University of Utah, Salt Lake City, Utah, December 1978.

(4) T. W. Anderson, "Estimation for Autoregressive Moving Average Models in the Time and Frequency Domain", *Annals of Statistics*, Vol. 5, No. 5, pp. 842-865:1977.

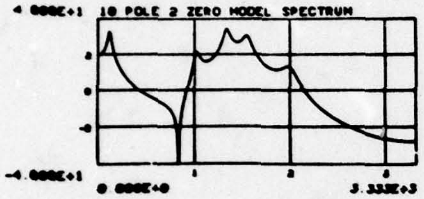


Fig. 1. 10-pole, 2 zero spectrum to be identified.

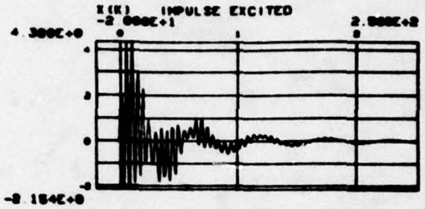


Fig. 2. System output when excited by impulse.

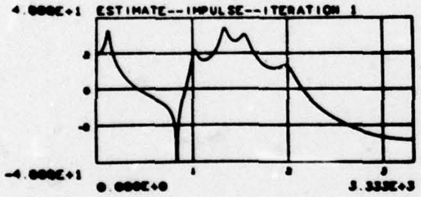
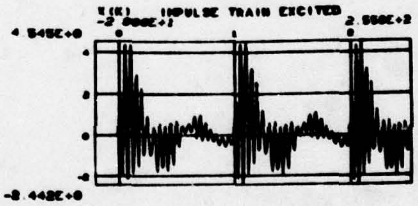


Fig. 3. Estimate of model spectrum from impulse-excited output.

Fig. 4. System output when excited by impulse train.



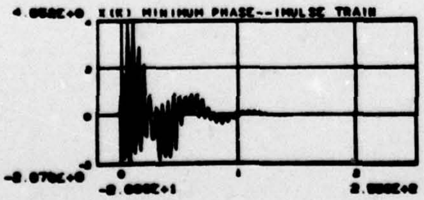


Fig. 5. Modified system output after cepstral processing.

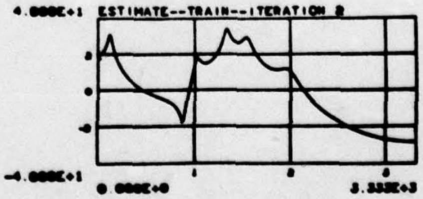


Fig. 6. Estimate of model spectrum from modified impulse-train-excited output.

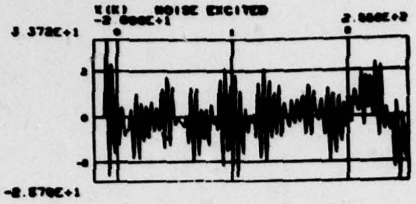


Fig. 7. Noise-excited output.

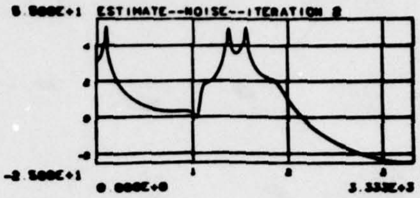


Fig. 8. Estimate of model spectrum from noise-excited output.

Table 1. Comparison of the Steiglitz and NR estimates for a(1) of an AR(1) process.

Iteration	Steiglitz	NR
1	-.05109	.49538
2	1.41350	.49538
3	-1.22940	.49538
4	-5.58270	.49538
5	-.24169	.49538
6	.89261	.49538
7	.97296	.49538
8	.99496	.49538
9	.99732	.49538
10	.99728	.49538

RATE/PITCH MODIFICATION USING THE CONSTANT-Q TRANSFORM

James E. Youngberg

December 1978

To be presented at ICASSP-79
April 2-4, 1979
Washington, D.C.

RATE/PITCH MODIFICATION USING THE CONSTANT-Q TRANSFORM

James E. Youngberg
Department of Computer Science, University of Utah
and
Soundstream, Inc.
Salt Lake City, Utah

ABSTRACT

Modification of the rate of occurrence of acoustic events without altering frequency content, and modification of pitch without changing time scale are presented as equivalent problems. While the short-time Fourier transform has been used to solve the rate modification problem, it is not a natural tool. It lacks the scaling property of the Fourier transform. The constant-Q transform, on the other hand, exhibits this property. A more natural rate/pitch modification system using the constant-Q transform is presented which performs well with rate/pitch changes by factors of between one-third and three.

INTRODUCTION

The possibility of modifying the rate at which speech is articulated has prompted a variety of efforts, ranging from simple time-base scaling and the excision or insertion of waveform segments to the more successful and complete approach used recently by Portnoff [1] involving time scaling of the short-time spectrum. These efforts have been hindered, in part, by the difficulty in satisfactorily defining what is meant by rate-changed speech. As Portnoff points out, this difficulty results from the ambiguity in classifying aural events as having either temporal or harmonic significance. Clearly, aspects of a signal which appear to manifest themselves harmonically should be preserved in rate compression or expansion of a signal, while characteristics which are perceived in time ought to be subject to modification. The short-time Fourier transform, which maps signals into a two-dimensional space, separates the time and frequency content of a signal based on the assumption that variations which occur over intervals greater than the constant time resolution of the analysis window will be classified as temporal events, while other characteristics must be measured along the frequency axis. Hence, while constant bandwidth analysis has produced very good speech compression and expansion results, it requires the making of signal-dependent assumptions about the time-frequency boundary. Evidence suggests that a constant bandwidth analysis criterion is consistent with neither the human auditory system in general nor with correct formulation of the rate compression/expansion problem in particular. For example, recent automatic phoneme recognition work by Searle [2] suggests that information by which various burst and stop phonemes are recognized occurs with time resolutions finer than 20 msec, and probably as fine as 5 to 10 msec. The auditory system, on the other hand, hears tones with fundamentals longer than 20 msec. Of course, the reason that the ear perceives stops and bursts as temporal events, while correctly analyzing 50 hz tones is that it is not a constant bandwidth device, but rather, constant-Q. Thus, the constant-Q transform, which maps signals into a two-dimensional space where time and frequency resolutions are dependent on analysis frequency, provides a more natural tool for performing independent modifications to temporal or harmonic aspects of signals. The problem, then, of defining what portions of a signal ought to be compressed or expanded in a speech rate change system is at least partially solved by requiring the time-frequency boundary to be a variable related to the ear's frequency-dependent boundary.

CONSTANT-Q ANALYSIS AND SYNTHESIS

The continuous formulation of a Fourier-like transformation which has a frequency-varying time-frequency boundary is given as

$$F(\omega,t) = \int_{-\infty}^{\infty} f(\tau)h((t-\tau)\omega)e^{-j\omega\tau}\, d\tau \qquad (1)$$

where the analysis window function, h, is defined to have finite non-zero extent (as in the Hann, Hamming, Bartlet and Blackman windows, for example). If by the single-argument functions, F and H, the Fourier integral transforms of f and h are designated, this constant-Q transform may be written in the form,

$$F(\omega, \nu-\omega) = F(\omega)H((\nu-\omega)/\omega)/|\omega| \qquad (2)$$

where ν is the frequency variable of the Fourier integral. Clearly, the analysis behaves for any analysis frequency, ω , as a bandpass filter whose frequency resolution is directly proportional to its center frequency, ω , whose time resolution is inversely proportional to ω , and whose output is frequency-shifted to zero. Because the ratio of analysis frequency to frequency resolution is a constant, this integral transform has been referred to as a constant-Q transform. By appropriately choosing the width and the shape of the analysis window, h, the time and frequency resolution of the constant-Q transform can be made to vary in a way which closely resembles the analysis performed by the human ear.

Although the time-frequency separation of data in the constant-Q spectral domain can be made to simulate that in the ear-domain, (1) is not useful in performing independent frequency or time scaling without a corresponding synthesis expression. The filterbank analogy to constant-Q analysis, explained above, suggests a method of synthesis. If the spectral domain is thought of as equivalent to the output of a contiguous bank of shifted, scaled lowpass filters, the possibility that a simple recombination of the various bandpass signals will lead back to the original signal seems obvious. The correct synthesis expression is, in fact, such an algorithm.

$$f(t) = k \int_{-\infty}^{\infty} F(\omega,t)e^{j\omega t}\, d\omega \qquad (3)$$

In this expression k is a constant which is determined by the nature of the analysis window, and which, in practice, is most easily determined empirically. That this analog to constant bandwidth FBS synthesis is but one of a family of possible synthesis forms has been pointed out by Kajiya [3].

SAMPLING THE CONSTANT-Q SPECTRAL DOMAIN

Implementation of (1) and (3) on a digital computer requires the proper sampling of the spectral domain. Schemes by which a constant bandwidth spectral domain may be sampled without loss of information are limited by the analysis window, which imposes time and frequency resolutions on the spectral information. Specifically, these time and frequency resolutions may be defined as the respective intervals over which the window function and its Fourier transform are "significant". The ambiguity in this definition is removed by defining the time resolution, T_∞, as the non-zero extent of h, and the frequency resolution, F_∞, as the extent of the principal interval around zero where H is positive. The scaling property of the Fourier integral transform guarantees that the product of the time and frequency resolutions will be a constant. Thus,

$$B_\infty = F_\infty T_\infty \qquad (4)$$

where B_∞ is a constant whose value is a consequence of the choice of the window function, h, and of the definitions of F_∞ and T_∞. Combined with the Nyquist theorem, this information is sufficient to permit sampling of the constant bandwidth spectral domain without loss of information. In particular, the density of time samples must be greater than F_∞, and the density of frequency samples must be greater than T_∞.

The extension of the above to the problem of sampling the constant-Q spectral domain is complicated by the dependence of T_∞ and F_∞ on frequency. The solution to this problem is enabled by the following formalization. First, define F_3 to be the more-limited principal extent over which H has values within 3 db of its maximum value, and let B_3 represent the constant product of T_∞ and F_3.

$$B_3 = F_3 T_\infty \qquad (5)$$

Then

$$Q = f/F_3 \qquad (6)$$

From this the time and frequency resolution may be determined as

$$T_\infty(f) = B_3 Q/f \qquad (7)$$

and

$$F_\infty(f) = B_\infty f/B_3 Q \qquad (8)$$

Equations 7 and 8, combined with the Nyquist theorem, give rise to lower bounds on the instantaneous sampling densities along the frequency and time axes respectively. In general, if $\Delta t(f)$ and $\Delta f(f)$ are the instantaneous sampling intervals at a frequency, f, then

$$\Delta t(f) \leq F_\infty^{-1}(f) \qquad (9)$$

and

$$\Delta f(f) \leq T_\infty^{-1}(f) \qquad (10)$$

Utilizing these limits and the filterbank formulation of the constant-Q transform suggested in (2), constant-Q analysis can be implemented in discrete form using fast convolution .

TEMPORAL AND HARMONIC SCALING

The approach to rate changes taken in the work reported here utilizes a property of the constant-Q transform not shared by the short-time Fourier transform. This property, the time scaling property, follows directly from (1) by substituting a time-scaled function, f(at), for f(t) and renaming the result.

$$F_1(\omega,t) = \int_{-\infty}^{\infty} f(a\tau)h((t-\tau)\omega)e^{-j\omega\tau}\, d\tau \quad (11)$$

Then, with a change of variables,

$$F_1(\omega,t) = \int_{-\infty}^{\infty} f(\tau)h((at-\tau)\omega/a)e^{-j\omega\tau/a}\, d\tau/|a| \quad (12)$$

or

$$F_1(\omega,t) = F(\omega/a,at)/|a| \quad (13)$$

Thus, the constant-Q time scaling property, given the relationship of (1), is

$$f(at) \overset{CQT}{\longleftrightarrow} F(\omega/a,at)/|a| \quad (14)$$

This property can be used to relate a change of scale of either the temporal or the harmonic spectral information to a change of scale of both the time domain signal and the other spectral axis. Assume, for example, the possibility of scaling the temporal axis of the constant-Q spectrum by a. This would give

$$F_2(\omega,t) = F(\omega,at) \quad (15)$$

If the signal, $f_2(t)$, resulting *from* substituting $F_2(\omega,t)$ into (3) were time scaled by 1/a, the result, using the constant-Q time scaling property would be

$$F_3(\omega,t) = |a|F_2(a\omega,t/a) \overset{CQT}{\longleftrightarrow} f_2(t/a) \quad (16)$$

$$F_3(\omega,t) = |a|F(a\omega,t) \quad (17)$$

Thus, as illustrated in figure 1, a harmonically scaled constant-Q spectral domain is related to a temporally scaled constant-Q spectral domain by a change of the signal's time scale.

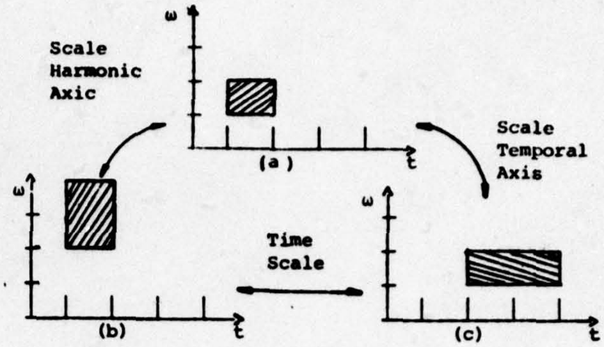


Figure 1. The relationship of axis scaling in the time and spectral domains (a) Original spectral event (b) Harmonically-scaled event (c) Temporally scaled event.

Because of the above duality, and because scaling of the harmonic axis is conceptually straight-forward, this approach was utilized in the work reported here to enable changes in either domain. Schematically, the harmonic scaler is implemented channel-by-channel as shown in figure 2.

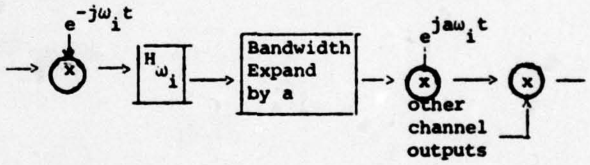


Figure 2. A single channel of a harmonic scaler. Channel center frequency is ω_i, scale factor is a.

Scaling of the harmonic axis by a factor, a, in a discrete implementation requires two operations. First, the effective bandwidth of each channel must be scaled so that the relationships among the various bands are not altered by scaling their center frequencies. Second, the center frequencies of the various channels must be scaled. This latter operation and the synthesis remodulation may be combined into an equivalent single operation. It should be noted, however, that to avoid aliasing during harmonic expansion (a>1), analysis channels must be adequately oversampled (by a factor of at least a). The more involved of the two operations above is the bandwidth scaling. Because the effective bandwidth of each channel is not directly proportional to the phase derivative as often assumed, the bandwidth of the channel is not accurately expanded by a when the phase derivative is scaled by a. Kahn and Thomas [4] have pointed out that the bandwidth of a simultaneously amplitude and phase modulated sinusoid is, in fact, a function of both modulating functions. Represent the modulated output sinusoid of the channel centered at ω_i as

$$c(t) = m(t)\cos[\omega_i t+p(t)] \qquad (18)$$

where m(t) and p(t), denoted below as m and p, are the (real) channel amplitude and phase functions. Then, for non-deterministic signals the channel bandwidth, Ω_c, can be estimated using the Kahn and Thomas bandwidth as

$$\Omega_c^2 = (E\{\dot{m}^2\}+E\{\dot{p}^2 m^2\})/E\{m^2\} \qquad (19)$$

where the E denotes the expectation and the dot the time derivative. Clearly, if the amplitude-modulating function is a constant, the approximation mentioned above is accurate. If, however, the amplitude portion of the bandwidth is a significant, but not dominant, portion of the total bandwidth, simple phase derivative scaling will lead to inaccurately scaled bands. This error can produce synthesized signals which exhibit reverberant effects reminiscent of comb filtering. Such effects are particularly evident in harmonically expanded signals (i.e. for a>1). To determine the a corrected factor by which the phase derivative should be scaled, assume that a correct bandwidth-expanded signal, c_a is given by

$$c_a(t) = m_a(t)\cos[\omega_i t+p_a(t)] \qquad (20)$$

and that

$$m_a(t) = a_1 m(t) \qquad (21)$$

$$p_a(t) = a_2 p(t) \qquad (22)$$

where a_1 and a_2 must be real. Then

$$a^2\Omega_{c_a}^2 = \; = (E\{\dot{m}^2\}+a_2 E\{\dot{p}^2 m^2\})/E\{m^2\} \qquad (23)$$

Note that a_1 has no effect. Substituting the equation (19) expression for Ω_c^2 into (23), the value of a can be determined.

$$a_2 = a[1+(1-a^{-2})E\{\dot{m}^2\}/E\{\dot{p}^2 m^2\}]^{\frac{1}{2}} \qquad (24)$$

This equation implies a conditional relationship between a and a_2. When expanding bands (a>1), the number by which the phase derivative must be scaled to scale the bandwidth by a is greater than a. When contracting bands (a<1), the opposite is true. It should be noted that when the amplitude modulation contribution, $E\{\dot{m}^2\}/E\{m^2\}$, dominates the total bandwidth, (24) may become imaginary. This effect corresponds to a lower limit on the total bandwidth reduction available using the assumptions of equations (21) and (22).

CONCLUSION

The constant-Q transform has been utilized to formulate a perception-related definition of the rate-change problem. This definition, along with the time scaling property of the constant-Q transform suggests a natural way of implementing both harmonic and temporal scale changes. The author's work proves this method to be capable of good quality compression and expansion for factors between one-third and three.

ACKNOWLEDGMENT

Thanks to James T. Kajiya, Tracy L. Petersen and H. Ravindra for their valuable insights and suggestions.

This research was supported, in part, by the Information Processing Branch of the Defense Advanced Research Projects Agency monitored by the Naval Research Laboratory under contract No. N00173-77-C-0041.

REFERENCES

[1] M.R. Portnoff, "Time-Scale Modification of Speech Based on Short-Time Fourier Analysis," Ph.D. Thesis, Dept. of Elec. Eng. and Comp. Sci., M.I.T., 1978.

[2] C.L. Searle, et. al., "Phoneme Recognition Based on Human Audition," Unpublished manuscript dated Oct. 1977. The principal author is with Queen's University, Ontario.

[3] J.T. Kajiya, "Toward a Mathematical Theory of Perception," Ph.D. Thesis, Dept. of Comp. Sci., Univ. of Utah, 1978.

[4] R.E. Kahn and J.B. Thomas, "Some Bandwidth Properties of Simultaneous Amplitude and Angle Modulation," IEEE Trans. on Inf.Theory, vol. IT-11. No. 4, pp. 516-520.